2016

Human Mutation/substitution Rate: Variability, Modeling And Applications

Varun Aggarwala
University of Pennsylvania, varunaggarwala01@gmail.com

Human Mutation/substitution Rate: Variability, Modeling And Applications

Abstract

Mutation generates genetic variation, and in turn selection purges deleterious variants from the population. Understanding both is critical for discovering causal genes and variants behind diseases or making inferences about evolutionary processes. Human mutation rate varies significantly across the genome although most studies have only considered the immediate flanking nucleotides around the polymorphic site to model and study patterns of variability. The impact of larger sequence context has not been fully clarified, even though it substantially influences rates of mutation. In the first part of this thesis, I develop a novel statistical framework and using data from the 1000 Genomes project, demonstrate that a larger heptanucleotide sequence context explains >81% variability in substitution probabilities, discovering novel mutation promoting motifs at ApT dinucleotides, CAAT, and TACG sequences. My approach also reveals previously undocumented variability in C-to-T substitutions at CpG sites, not immediately explained by differential methylation intensity. Building on this framework, I model the selective forces acting on the coding genome and develop statistical scores that measures the intolerance at the gene or amino-acid level for functional variants. I demonstrate clinical utility of such intolerance scores in identifying genes associated with multiple human diseases including Autism. Next, I apply these lessons of mutation rate variability to develop an algorithm to detect sub-genic enrichment of de novo germline mutations in RB1 gene of bilateral Retinoblastoma (RB) probands to further elucidate disease biology. I demonstrate that previously noted 'hotspots' of nonsense mutations in RB1 are compatible with the elevated mutation rates expected at CpG sites, refuting a specific mechanism in RB pathogenesis. I also find enrichment of splice-site donor mutations of exon 6 and 12 but depletion at exon 5, indicative of previously unappreciated heterogeneity in penetrance within this class of substitution. Finally, I generate more accurate and informative estimates of de novo germline mutation rate in humans, and develop a toolkit to simulate, distribute and interpret mutations in human diseases. Overall, my research uncovers novel variability in human mutation rate and provides a systematic framework for analyzing mutational data, which can be used from causal gene discovery to elucidating specific disease mechanisms.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

Benjamin F. Voight

Keywords

Computational Biology, Evolutionary Biology, Human Genetics, Mutation Rate, Statistical Genetics

Subject Categories

Bioinformatics | Genetics

HUMAN MUTATION/SUBSTITUTION RATE: VARIABILITY, MODELING AND APPLICATIONS

Varun Aggarwala

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

_____

Benjamin F. Voight, Ph.D., Assistant Professor of Pharmacology and Genetics

Graduate Group Chairperson

_____

Li-San Wang, Ph.D., Associate Professor of Pathology and Laboratory Medicine

Dissertation Committee

Struan F.A., Grant, Ph.D., Associate Professor of Pediatrics

Warren J. Ewens, Ph.D., Emeritus Professor of Biology

Christopher D. Brown, Assistant Professor of Genetics

Suzanne M. Leal, Professor of Molecular and Human Genetics, Baylor College of Medicine

John B. Hogenesch, Professor of Molecular and Cellular Physiology, University of Cincinnati

HUMAN MUTATION/SUBSTITUTION RATE: VARIABILITY, MODELING AND APPLICATIONS

DEDICATION

To my best friend and my partner in life, my dear wife Ankita,

You are my inspiration, my strength, and the joy of my life.

To my great nation India, in whose service I want to dedicate the rest of my life.

ACKNOWLEDGMENT

Whenever I come across a thesis, I always make it a point to read the acknowledgement section. While the thesis abstract tells the reader about the beautiful scientific journey, the personal acknowledgement section sheds light on the unique and interesting life story of the author. Now, it is my chance to pour my heart out and acknowledge everyone who has helped me reach this stage of my career.

First and foremost, I want to thank my thesis advisor Ben Voight for helping me mature from a young naïve student to a scientist who is full of ideas and promise. Like any other relationship, we also had our ups and downs, but deep inside I have nothing but gratitude for him. Thank you Ben for teaching me how to read, write, speak and think like a scientist!

Next, I want to thank my committee members (Struan Grant, Casey Brown, Warren Ewens, Suzanne Leal and John Hogenesch) for all the advice and help throughout my graduate studies. I also want to thank our collaborator Arupa Ganguly for the Retinoblastoma dataset, whose analysis using my mutation rate framework, forms the bulk of Chapter 4.

I am indebted to Warren Ewens for teaching me everything about statistics. Warren's class was the single most important event in my graduate career. It improved my confidence and helped me realize that I can have a successful career doing research which heavily uses statistics.

Special thanks are due to John Hogenesch for being a superhero and for everything, and literally everything. John's GCB 531 was the first biology class, that I took after junior high school. I could not have chosen a better class. It ignited my interest in genomics and computational biology. Thank you John for helping me get into the GCB program for my PhD studies. Thank you for suggesting me to work with Ben for my thesis. Thank you for giving me the best advice at all stages of my career. And, most importantly, thank you for teaching me how to be a brave and a fearless scientist.

Next, I want to thank the GCB program for admitting me when no one even wanted to touch me with a 10-foot pole. Those were difficult times and I am incredibly grateful to the GCB graduate

group for providing me with a home. While I am a vocal critic of our program policies, I only do that because I truly and deeply care about it. Deep inside, I think I could not have chosen a better home to conduct my graduate studies. Thank you Maja Bucan and now Li-San Wang for being terrific graduate chairs. You truly care about the students, and this distinguishes you from everyone else. Thank you Hannah Chervitz and Maureen Kirsch for being so capable, quick and efficient administrators. Thanks you Maureen for always being so positive and caring!

Next, I want to thank Li-San Wang for helping me transition from my dark days in computer science to my bright and fun filled days in the GCB program. I cannot imagine the trajectory of my career if I would not have spoken to you in May 2010. Thank you for being so wise and calm.

I also want to thank the Penn CIS department for giving me a master's degree on the way and letting me pursue my interests in genomics and computational biology. Thank you Jianbo Shi and Lyle Ungar for all your support during my very difficult phase in 2009-10. Thank you CMU CS department for admitting me in 2008 and for providing me my very first home in USA. While I ultimately moved on from just focusing on computer science, my prior experiences have helped me tremendously in discovering both myself and my interests in the broad field of genomics and computational biology.

I am sure I have forgotten to thank many amazing people who have both directly and indirectly helped me along the way. Trust me, it actually took a big village to raise me and help me finish my PhD studies. I wish I can also give back to the amazing community of professors, scientists and fellow graduate students.

Next, I want to thank all my friends and well-wishers who have nurtured me in my graduate school life. Specials thanks to the best roommates ever: Pavan Nukala and Mukund Raghothaman. You are both wonderful human beings and without your friendship and tolerance, I would have broken down completely and would have gone back to India. Thank you for not making me miss my family so much. Special thanks to my friends Krishna Vijayendran, Vince Luczak, Onur Yoruk and Christel Chehoud (Sjoland). Krishna, I deeply cherish our state of the

union talks every month. I hope we can have them for the rest of our lives. Vince, thank you for our very regular and detailed political discussions. Onur, you are a true gentleman and the nicest person in this world. Christel, you are an ultimate go-getter and a wonderful friend. I also want to thank the members of Voight Lab with whom I have had countless discussions on everything. Special shout out to Paul Babb who helped me a lot in improving my communication skills and is also the best lab-mate ever! Many thanks to “Lord” Nick Lahens for being a source of wisdom and fountain of calmness. Moreover, I want to thank my other non-Penn affiliated friends Mayank Maheshwari, Govind Kothari, Ankit Kumar, Paritosh Chaube and Mayuri Rege (who is now at Penn) for their support and all the friendly distractions. If everything around me fails, I will still consider myself incredibly lucky because of these amazing bonds of friendship that I have forged in my graduate school days. Again, I have missed a lot of names here. Trust me, you still have my deepest gratitude. Your act of kindness and friendship has touched me in more ways than I can describe.

Now, I want to thank my family members for all the love and support throughout this phase of my life. Thank you mummy and papa, for being the best parents ever. I am still your little boy, and my day is only complete after our daily skype calls at 10:30 PM EST, and when it is morning in India. Thank you *bhaiya* (elder brother) and *bhabhi* (sister in law) for our weekly skype calls and for always checking up on me. I could not have wished for a more perfect family. We will always be together even if all of us are in different countries and time zones! I feel truly fortunate because now I have an extra set of equally caring and loving parents through my marriage. Thank you for loving me so much and also for being more fun than my *biological* set of parents!

Last but not the least, I want to thank my dear wife Ankita for everything. I tend to exaggerate at times, but words are not enough to express my deep feelings for you. YOU ARE SIMPLY THE BEST. Thank you for making my life complete in all ways possible. Thank you for making me the luckiest person on earth and for giving me very soon, a more perfect yet innocent version of both of us together, our daughter Nysa!

ABSTRACT

HUMAN MUTATION/SUBSTITUTION RATE: VARIABILITY, MODELING AND APPLICATIONS

Varun Aggarwala

Benjamin F. Voight

Mutation generates genetic variation, and in turn selection purges deleterious variants from the population. Understanding both is critical for discovering causal genes and variants behind diseases or making inferences about evolutionary processes. Human mutation rate varies significantly across the genome although most studies have only considered the immediate flanking nucleotides around the polymorphic site to model and study patterns of variability. The impact of larger sequence context has not been fully clarified, even though it substantially influences rates of mutation. In the first part of this thesis, I develop a novel statistical framework and using data from the 1000 Genomes project, demonstrate that a larger heptanucleotide sequence context explains >81% variability in substitution probabilities, discovering novel mutation promoting motifs at ApT dinucleotides, CAAT, and TACG sequences. My approach also reveals previously undocumented variability in C-to-T substitutions at CpG sites, not immediately explained by differential methylation intensity. Building on this framework, I model the selective forces acting on the coding genome and develop statistical scores that measures the intolerance at the gene or amino-acid level for functional variants. I demonstrate clinical utility of such intolerance scores in identifying genes associated with multiple human diseases including Autism. Next, I apply these lessons of mutation rate variability to develop an algorithm to detect sub-genic enrichment of *de novo* germline mutations in *RB1* gene of bilateral Retinoblastoma (RB) probands to further elucidate disease biology. I demonstrate that previously noted 'hotspots' of nonsense mutations in RB1 are compatible with the elevated mutation rates expected at CpG sites, refuting a specific mechanism in RB pathogenesis. I also find enrichment of splice-site donor mutations of exon 6 and 12 but depletion at exon 5, indicative of previously unappreciated

heterogeneity in penetrance within this class of substitution. Finally, I generate more accurate and informative estimates of *de novo* germline mutation rate in humans, and develop a toolkit to simulate, distribute and interpret mutations in human diseases. Overall, my research uncovers novel variability in human mutation rate and provides a systematic framework for analyzing mutational data, which can be used from causal gene discovery to elucidating specific disease mechanisms.

TABLE OF CONTENTS

CHAPTER 1: Introduction

Human mutation rate

Mutation is the most important force that has shaped our genomes since we evolved from single cell species [1]. It generates genetic variation on which evolution acts [2], causes variation between individuals of same [3] and different species [4], results in cell to cell heterogeneity [5] and is responsible for genetic disorders; inherited [6] or somatic [7] like cancer. Hence, understanding how mutations originate across the genome is fundamental to our understanding of life. Previous work over the last century, has successfully discovered and described the process of mutation [3, 8] in single celled organisms to complex eukaryotic species like *Homo sapiens*. Here, I define mutation rate as the measure of the rate at which mutations occur in the human genome over time. An accurate estimate of the human mutation rate can shed light on the process of mutagenesis and is essential for all aforementioned quantitative applications which model and utilize the properties of mutation.

There are three major approaches to study the human mutation rate. The earliest is the phylogenetic approach [3, 8–10], which measures fixed substitutions at neutral sites between species to find an estimate of mutation rate. Traditionally, fixed substitutions at putatively neutral evolving pseudogenes were measured to get an estimate of mutation rate. This approach also requires the divergence time between species, which is estimated from fossil dating [11] to get an estimate of mutation rate. This approach has been quite informative and provided the initial clues about the general properties of mutation and its variation, which helped improved our understanding of the process of mutagenesis [3, 8]. However several issues remain with this approach as it makes several assumptions about the accuracy of species divergence time, and putatively neutral evolving nature of pseudogenes [12]. Moreover, issues of multiple substitutions at a site and small number of substitutions at pseudogenes also add uncertainty to the mutation rate estimates. Next, with the advent of next generation sequencing the community has tried to

infer the properties of mutation from large scale population sequencing datasets [13, 14]. Assuming that intergenic noncoding regions are neutrally evolving and hence effects of selection are minimized, they can be studied to understand the properties of mutation rate [11, 15]. This approach is informative, however we only learn about general properties of mutation and how it has shaped the genetic variation spectrum in putative neutral regions, but we do not get actual estimates of mutation rate. Another informative approach involves resequencing of family trios (*i.e.*, parents and the offspring) to identify *de novo* mutations, from which a rate per generation can be estimated [16–18]. This approach is unbiased, but suffers from issues of extremely small sample sizes as only ~70 *de novo* mutations are observed genome-wide in a family. Even with larger family sizes, resulting in thousands of mutations, we are limited in our power to completely understand the properties of mutation and accurately infer the mutation rate. Moreover, due to numerous filters and quality control measures, many real mutations are omitted from the final set of high quality *de novo* mutations. In future, with large sample sizes and higher fold coverage of sequencing studies these issues can be resolved, but in current form the direct sequencing approach is severely underpowered. Hence, despite statistical and technological advances, many gaps in knowledge about the process of mutagenesis, related to mechanism and discovery of novel determinants of mutation still remain.

Mutations can arise due to both endogenous or exogenous sources. Endogenous sources might be due to errors introduced during replication [19, 20] or due to spontaneous changes [21, 22]. Well characterized sources of exogenous factors resulting in mutation are radiation and chemical agents [23]. While most errors are corrected by the DNA repair machinery [24], if left uncorrected it results in a mutation. Since replication introduced errors result in mutation, this also explains the high paternal effect on human mutation rate due to the increased number of cell divisions in spermatogenesis [25]. Most replication errors arise from incorrect base incorporation resulting in point mutations or slippage of polymerase which results in short indels [26]. It is important to note that DNA repair is more effective in regions where DNA strands can be readily separated [27].

Since, A-T base pairs have two hydrogen bonds, in contrast to the three stronger hydrogen bonds at G-C base pairs, it is more difficult to separate DNA in GC rich region, resulting in a higher rate of mutation here. An exquisite example of spontaneous change resulting in a higher mutation rate is the deamination of methylated cytosine at CpG sequence context [28, 29]. Multiple studies have characterized the higher mutation rate here (~15 fold higher than the genome-wide average) and this context is almost exclusively modeled in all studies of mutation rate and its variation [30–33]. Well established exogenous factors that induce mutation includes UV light exposure which can result in C-to-T transition at di-pyrimidine sequences [34] or G-to-T transversions caused by Benzo(a)pyrene [35]. Characterization and understanding of mutational signatures and associated mechanism has direct relevance in multiple human diseases including cancer. The advent of next generation sequencing has revolutionized the field of mutation research, as novel mutational signatures [36] are observed from direct sequencing of germline and somatic tissues (mostly in different cancer). However, many open questions about the process of mutagenesis still remain and systematic approaches are needed to discover novel mutational signatures from large scale sequencing datasets, which can then elucidate the associated mechanisms and factors causing mutation.

Since mutations arise due to a host of different exogenous and endogenous factors, which result in different mutational signatures, it is natural to expect variation in the genome-wide rate of mutation. Previous studies over the last few decades have reported extensive variation in the mutation rate across the genome [3, 8], much more than one would have expected from known sources which confer variability. Mutation rate has been shown to vary at a broad scale of chromosome to individual base pairs. Over broad scales, replication timing has been shown to influence the rate of mutation. In particular, the late replicating regions [19, 20] have been shown to have a higher mutation rate and associated SNP density, potentially due to scarcity of free nucleotides during replication. Processes like transcription coupled repair, has also been shown to affect the rate of mutation [37]. Since this process is more prominent over actively transcribed

regions, more mutations are observed in the non-transcribed strand and this asymmetry is more pronounced for highly expressed ubiquitous genes [38]. At fine scale, local sequence context has been shown to strongly influence the rate of mutation, much more than broad scale effects. First, we observe a two-fold higher rate of transitions than transversions, which can be explained by relative ease of purine to purine or pyrimidine to pyrimidine DNA mispairing during replication, resulting in a transition mutation [39]. Second, adjacent sequence context explains significant variation in the rate of mutation [30]. For example, it captures properties of spontaneous deamination of methylated cytosine resulting in a thymine mutation at CpG motif. Virtually all studies use trinucleotide sequence context (which considers one nucleotide at either 5' and 3' end) around a polymorphic site to summarize and model the genome-wide variability in mutation rate from applications ranging from basic science understanding of mutagenesis [32, 33], in different cancers, and in developing clinical applications [31]. However, open questions still remain. Does the trinucleotide sequence context capture all or most variation in mutation rate? If not, what size window of sequence context best explains patterns of genetic variation we observe in human populations? With improved models that explain observed data to a greater extent than previous possible, (i) are there features beyond CpG sites that further enumerate the processes and potential mechanism that cause mutation, and (ii) how can we employ this improved understanding to power studies of human disease? In this dissertation, I systematically address these questions related to mutation rate variability.

Purifying selection

Purifying or negative selection is the act of purging of deleterious variants from the population [40]. Since new mutations in functional regions like protein coding genes are mostly deleterious and reduce organismal fitness [41, 42], purifying selection by removing these mutations maintains the genomic integrity. Forces of evolution and drift over the course of time have optimized the genome of most species and hence new deleterious mutations are kept from increasing in number and taking over the fixed variants. It is also important to note that some new mutations

can also increase fitness, hence they are selected for in the population and reach fixation [43]; this phenomenon is called positive selection. Moreover, in cases where having heterozygous alleles is beneficial, evolution acts to actively maintain multiple alleles in the population [44]; this phenomenon is called balancing selection. While both positive and balancing selection has been observed in the genome, purifying selection remains the most common and prevalent of all [45].

Purifying selection is strongest on variants in the protein coding genes or regions with regulatory potential, because they have a higher likelihood to adversely affect organismal fitness [46]. Purging of deleterious variants can also remove linked variation [40], resulting in a reduction of variation surrounding the locus under selection; this phenomenon is called background selection.

The signal of purifying selection is used to identify functionally important and highly conserved regions from comparative genomic approaches, because it manifests as lack of variation at a locus. Improvement our understanding of the frequency, occurrence, and strength of negative selection can inform the genetics of both human disorders and predict population frequencies of causal disease alleles which can further aid in screening and surveillance [47]. We expect strong purifying selection on genes associated with monogenic disorders because highly penetrant alleles in these genes can severely reduce organismal fitness [47, 48]. In contrast, weaker purifying selection may act on alleles in many genes which individually confer some risk for a complex disease [49]. Overall, genes evolving under weak purifying selection or constraint are thought to be less important or have more redundancy than those evolving under strong purifying selection [50].

Existing approaches to discover causal variants and genes for different disorders often utilize some form of conservation or purifying selection in their methodology [51]. The community has successfully identified numerous causal genes for both monogenic and complex diseases by identifying regions undergoing strong purifying selection and then prioritizing those regions for follow up [47, 50, 52, 53]. Such approaches will vastly benefit from more accurate estimates of

purifying selection and its strength of action on the genome. Approaches for finding purifying selection have largely focused on comparative genomic approaches to find evolutionary conserved regions [54, 55] or by considering the site frequency spectrum and measuring the enrichment of low frequency and reduction in medium and high frequency variants undergoing strong selection [56]. While these approaches have been informative, open questions and challenges still remain. Have we identified the force of selection and constraint on all genes and sub-genic locus? Can we further improve the estimates of selection and rank regions by strength of selection on them?

It is important to note that human genetic variation is shaped both due to forces of mutation and selection [57]. With an accurate model for the background rate of mutation, it should be possible to characterize the effects of background selection, based on the difference between the genetic variation compared to the model expectation at a locus (or gene). While this does not result in a *direct* estimate of purifying selection, different loci can be compared to each other and can be ranked by the deviation from this expectation as a proxy for the strength of negative selection at the gene. In this dissertation, I model the rate of mutation across the genome and, using those rates, obtain an estimate of purifying selection at each translated gene in the genome, which can then be used in several clinical applications.

Landscape of genetic diseases

Deleterious mutations in protein coding genes can cause several diseases [58]. These deleterious mutations can either be inherited, occur *de novo* or as somatic changes in the affected proband [59]. Adding an extra layer of complexity, genetic diseases have a spectrum of genetic architecture, ranging from single gene (monogenic) to multiple genes in combination with environmental factors (complex, multifactorial or polygenic). Here, I will review two diseases, one single gene and one complex (multi-gene).

Retinoblastoma, a monogenic disease

Retinoblastoma (RB) is a cancer of the developing retina caused due to bi-allelic inactivation of the tumor suppressor gene, *RB1* [60]. The incidence rate of RB in human population is approximately 1 in 20,000 live births [61]. It occurs in non-hereditary form in ~60% of cases, characterized by tumor in one eye and is also known as unilateral Retinoblastoma [62]. The non-hereditary form of RB occurs later in childhood and is not associated with a higher risk of other cancers. In contrast, the hereditary form of RB, with a frequent clinical manifestation of tumors in both the eyes, is characterized by germline mutations - either inherited or *de novo* - in the *RB1* gene [63]. This form of RB is also called bilateral Retinoblastoma and it occurs early in childhood, and is also characterized by higher risk of other cancers.

Knudson [64] proposed a two-hit disease theory for Retinoblastoma. According to this theory, RB is caused by mutations in both copies of the tumor suppressor *RB1* gene. If one mutation in the *RB1* gene is present in germline, then only one another somatic mutation is needed to develop RB. This explains why bilateral RB, with a germline mutation already present in *RB1* gene, occurs early in childhood and is characterized with cancer in both the eyes. However, unilateral Retinoblastoma with no germline mutations require two somatic mutations which cause inactivation of the *RB1* gene resulting in delayed onset with tumor in one eye.

Clinical sequencing of the *RB1* gene, management and counseling of RB is now a routine practice for patients and their families [65, 66]. Because individuals with bilateral form of Retinoblastoma can transmit the germline mutations to their offspring, it is necessary to distinguish this form of RB from unilateral cases. Moreover, individuals with bilateral RB are also more likely to develop second malignancies, so the clinical management of these individuals allow for appropriate treatment and long-term surveillance for other cancer types.

Multiple groups over the last few decades have sequenced the *RB1* gene in patients with both forms of Retinoblastoma and have hypothesized several disease mechanisms from large mutation datasets [67, 68]. Several pathogenic mutations and their role have been identified [69–

72], such as an abundance of nonsense mutations in RB patients which result in loss of function of the *RB1* protein, or patterns of penetrance and pathogenicity in missense mutations. However, several open questions about more pathogenic sub-genic locus in *RB1* gene which can further elucidate disease mechanism still remain. In this dissertation, I will use concepts from mutation rate variation to analyze RB mutation datasets along with large-scale surveys of population level variation in the coding genome [73] to test hypotheses about pathogenicity with sub-sequences of *RB1*.

Autism, a complex disease

Autism is neurodevelopmental disorder characterized by impairment in social interactions and repetitive behaviors [74]. It occurs in roughly 1% of the population, although the rate of incidence has been increasing in recent times [75], mostly due to increased awareness and broadening of diagnostic criteria. Males are roughly 4 times more likely than females to have a diagnosis of Autism [76]. Autism is a spectrum disorder and patients show vast phenotypic heterogeneity in cognitive and language abilities [77]. Moreover, Autism spectrum disorder rarely occurs in isolation, and often coexists with other neuropsychiatric and medical conditions like epilepsy, intellectual disability, anxiety, attention-deficit/hyperactive disorder and gastrointestinal problems [78]. Autism spectrum disorders (ASD) are generally diagnosed in early childhood before the age of 3 but symptoms often manifest late into adulthood.

ASD has a strong genetic component as identified from twin studies and recurrence risk in families [79]. Twin studies for ASD have demonstrated a concordance rate of ~90% in monozygotic and ~up to 20% in dizygotic twins [80, 81]. Epigenetic and environmental factors are also thought to be influence susceptibility to ASD, since concordance rate in monozygotic twins does not reach 100%. Early genetic causes of ASD were identified in monogenic diseases such as fragile X syndrome, neurofibromatosis and Rett syndrome [82, 83] and from large scale chromosomal duplication or deletion [84]. However, with improved genotyping and sequencing

technologies, pathogenic mutations were identified in several candidate genes and they together highlighted the role of synaptic dysfunction in ASD susceptibility [85]. Multiple genotyping studies have also demonstrated the role of small and large copy number variants, both inherited and *de novo* in individuals with ASD [86–88]. Recent sequencing studies have also identified a higher burden or rare and *de novo* variants in intolerant genes (under a strong selective constraint [31]), specifically encoding proteins for synaptic formation, transcriptional regulation and chromatin remodeling pathways [52, 89, 90]. However, many open questions related to discovery of causal genes for Autism which can explain a higher percentage of heritability [91], and associated mechanism still remain.

In this dissertation, I will use concepts from mutation rate variation to analyze a larger *de novo* sequencing dataset for Autism probands, and test for role of several causal genes in ASD pathogenesis.

Sequencing approaches

Population level sequencing

This approach involves sequencing of unrelated individuals in a population to find the polymorphic events present in the genome [13, 92]. In this dissertation, I mostly focused on single nucleotide polymorphisms (SNPs), though population level sequencing also detects indels and copy number variations. An important goal of population level sequencing involves finding the frequency of different polymorphisms, which can help understand the role of mutation, selection, drift and other evolutionary forces in shaping the human genome [15, 57, 93]. Various studies also perform population sequencing in both affected cases and controls, to understand the role of common and low frequency variants in disease pathogenesis [94, 95]. Studies which focus on understanding population genetics parameters, generally sequence the entire genome of the individuals [13], while disease based studies have mostly focused on exome sequencing [96] or targeted sequencing of genomic regions [97].

The choice of tissue for most population level sequencing studies is blood, because of ease of collection and quality of DNA. However, because of issues of mosaicism [98], many studies sequence multiple tissues for accurate detection of polymorphisms present in the germline. Since population level sequencing entails sequencing of unrelated individuals to find polymorphisms and the associated frequency of alleles in the population (*i.e.*, the site frequency spectrum), special emphasis is given on finding relatedness and detecting ancestry of individuals in a population [99].

Rapid sequencing of multiple individuals in a population has been made possible due to advent of next generation sequencing technology [100], and advances in computing infrastructure for storage of large datasets and increased computing power for sequence alignment and variant calling. Moreover, due to population level sequencing of unrelated individuals in a population, accurate variant calling from low pass sequencing (less coverage) is possible. However, some regions of the genome are hard to sequence [101] and hence variant calling is still a challenging task. Nonetheless, with high quality sequencing datasets and more coverage, the community is trying to successfully identify more polymorphisms and study them in different contexts.

In this dissertation, I will use population level sequencing data to learn more about properties of mutation rate variation, selection and will infer several population genetics quotient of interest.

Family based sequencing

This approach entails sequencing of biological parents [16, 17] and their offspring, to find *de novo* germline mutations in the offspring. Typically, the exomes are sequenced for the entire family [102], but whole genome sequencing has become common due to increasing availability and better pricing [18]. The procedure involves (a) sequencing the family trio (mother, father and the child), (b) calling the polymorphic variants in all of them separately, and finally (c) finding the variants present uniquely in the offspring. The unique variants present in the offspring, but not in the biological parents are the *de novo* germline mutations. Studies with large scale sequencing of

family trios have reported an average of ~72 *de novo* mutations present in each offspring after sequencing of entire genome [16]. This corresponds to a genome-wide mutation rate of $1.2*10^{-8}$ mutations, per nucleotide per generation. Similarly, this results in ~1 *de novo* mutation in each offspring after sequencing of entire protein coding genome [16, 90].

Typically, the blood is the DNA source of choice for most sequencing studies, because of ease of collection and quality of DNA. Hence, the *de novo* mutations identified in the offspring can also be caused by mosaicism [98], although with very less probability. However, for the purpose of analysis the community assumes it to be germline *de novo*. Moreover, the current strategy cannot distinguish recurrent from actual *de novo* mutations. Since, the probability of recurrent and the *de novo* mutation to happen at the same position is very low, the community considers all the new mutations in the offspring to be *de novo.*

While family based sequencing is mostly used for finding *de novo* mutations in the offspring, there are additional advantages to family-based studies. First, family based sequencing allows for more accurate variant calling because of sampling of four independent alleles at each position, a total of six times (2 from each parent, and 1 from child and 2 copies in each). Since each variant in child is inherited from parents, with the exception of *de novo* events, this allows for more accurate joint variant calling in families. Second, it improves the ability to call variants in low coverage areas of the genome. Since less reads are available for accurate calling of variants in low coverage areas, joint calling of variants with the offspring significantly improves the sensitivity of the method to call variants with no effect on specificity. However, joint calling of variants has a high bar (more high quality reads in the child at the mutated site) for a variant to be reported as *de novo*, so it might also result in under calling of *de novo* variants in a family. Third and finally, family based sequencing allows for observation of inheritance of variants. Since, both the copies of the genome are sequenced in the entire family trio, this makes possible assigning the parent of origin status to each variant.

The community has mostly utilized family based sequencing approach to understand the process of mutagenesis [16, 17], for example understanding the relationship between paternal age and *de novo* mutations, or to investigate the role of *de novo* mutations in several pediatric diseases like Autism [90], Retinoblastoma [67], or Epilepsy [103]. While the former mostly entails genome-wide sequencing, hence the dataset are small, the latter involves mostly exome sequencing of larger samples.

In this dissertation, I will utilize a family based sequencing dataset to validate my estimates for mutation rate variation. Moreover, I will use several *de novo* sequencing datasets of Autism and bilateral Retinoblastoma disease to find causal genes, testing specific, disease-relevant hypotheses.

Thesis outline

The central objective of this dissertation is to model the variability in the genome-wide rate of mutation and then leverage this variability to find causal gene and variants behind diseases, and test for specific disease hypotheses. A widely used yet unexplored idea in the field is, if sequence context models can explain variation in mutation rate, and furthermore do larger sequence context windows do better. Moreover, once the variability in mutation rate has been modeled, there is a lack of a systematic framework which can use this in several clinical applications. In this thesis, I address these open questions and challenges.

In chapter 2, I evaluate different sequence context models and conclude that a larger heptanucleotide sequence context (which considers three nucleotides on either 5' and 3' end) around a polymorphic site explains significantly more variability in nucleotide substitution probabilities, in contrast to the commonly used trinucleotide sequence context (which considers one nucleotide on either 5' and 3' end). I use millions of variants from 1000 genomes[13] project, to find the substitution probabilities and show that they capture features of germline *de novo* mutation rate, identify several novel mutation promoting motifs at TA<u>C</u>G, <u>A</u>T and CA<u>A</u>T sequence

contexts, and investigate the complex relationship between methylation intensity and higher CpG mutation rate.

In chapter 3, I build on the substitution probability framework, and estimate the strength of selection at each sequence context in the coding genome. I use these estimates of selection to prioritize pathogenic variants and develop intolerance scores at gene and amino acid level for functional changes. Finally, I demonstrate how the intolerance scores can be used to prioritize genes for follow up from an Autism *de novo* sequencing dataset.

In chapter 4, I develop a generalized approach to find enrichment of mutations at any sub-genic locus, beyond expected from a background model of mutation rate variation. I use this approach on a bilateral Retinoblastoma dataset to test for and discover regions in the *RB1* gene with an unusual enrichment of *de novo* germline mutations, suggesting additional pathogenicity. I find that nonsense mutations in *RB1* gene are enriched in RB ascertained probands and hence overall more pathogenic, but the previously identified higher number at CpG sites can be explained by higher background mutation rate, suggesting no additional pathogenicity at CpG sites compared to other nonsense mutations. Finally, I also test for and discover enrichment and hence more pathogenicity in specific splice-sites and missense mutations.

In chapter 5, I estimate the *de novo* mutation rate at different sequence contexts using the substitution rates in chapter 2. I demonstrate that heptanucleotide mutation rate estimates accurately and best explain patterns of *de novo* mutations in an external dataset, in contrast to the commonly used trinucleotide context based mutation rate estimates. Finally, I use these rates to develop a toolkit to simulate, distribute and interpret *de novo* mutations in human diseases.

CHAPTER 2: Sequence context models and variability in substitution/mutation rates

Introduction

Measured at the level of the chromosome down to the individual base, rates of single nucleotide substitution vary substantially by position across mammalian genomes, including the humans [3]. An exquisite example of the role for sequence context in contributing variability in substitution rate are CpG dinucleotides, where spontaneous deamination of 5-methylcytosine results in ~14 fold higher C-to-T substitution rates [21, 104]. Modeling the variability in nucleotide substitution rates will inform our understanding of evolutionary processes, help identify functional noncoding regions [105] and mutation promoting motifs, reveal mechanisms behind spontaneous mutation, and aid in prediction of the clinical impact of polymorphisms discovered through resequencing [96]. Such models will need to determine not only the optimal window of local sequence context, but should also integrate knowledge of functional constraint on the genome owing to pressure from purifying selection.

Studies of complex human disease have incorporated a simple trinucleotide sequence context [30, 106] into models to quantify the probability of *de novo* mutational events [33, 89], to clarify the distribution of somatic mutational events segregating in different cancers [32], and to model the purifying selective pressure on gene sequences [31]. As their focus was clinical, these reports did not determine if this context model best captured the extent to which flanking nucleotides impact the variability in genome-wide nucleotide substitution rates. Here, I report a statistical framework that compares the extent to which different local sequence lengths impact the probability of nucleotide substitution, tested using data from the 1000 Genomes (1KG) Project [13]. I define the probability of nucleotide substitution as the chance that a nucleotide in the human genome reference is polymorphic – i.e., the nucleotide position segregates alternative nucleotides within the population. This probability depends upon population history, selection,

sample ascertainment, and local context features that influence the rate of mutation. I show that a larger sequence context that considers 3 base pairs on either side of the polymorphic site (heptanucleotide sequence context) explains >81% variability in substitution and is significantly more informative than the commonly used trinucleotide sequence context. My approach discovers novel mutation promoting motifs at ApT dinucleotides, CAAT, and TACG sequences. I also identify previously undocumented variability in C-to-T substitutions at CpG sites, not immediately explained by differential methylation intensity. Finally, I show that the substitution probabilities, inferred over the intergenic noncoding region of the genome, capture features of germline *de novo* mutation rate.

Data access

Sourcing population samples

Samples were obtained from phase 1 of the 1000 Genomes Project. Further details about sample collection, sequencing, and variant calling are available in the original publication [13]. I considered only the variants from African (n= 246 individuals), European (n = 379), and East Asian (n = 286) ancestries.

Selection of intergenic non-coding sequences

Intergenic sequenc [107] (Ensembl Genes 75 and Homo sapiens genes GRCh37.p13) and RefSeq Genes [108]. I initially removed centromeric, telomeric, and repetitive regions from these non-coding sequences by filtering out the contiguous sequences at the ends of the chromosomes and “gene deserts” of length greater than 2 MB. I also filtered away the sequences that were not present in the combined accessibility mask (version 20120824) of the 1000 genomes project. As a result, I was left with ~1100 Mb of autosomal intergenic regions and ~90 Mb on the X chromosome. Within these intergenic regions, I found 10,809,273 variants in the African

populations, 7,051,667 variants in the European populations, and 6,024,240 variants in the East Asian populations.

Selection of HapMap variants

Single nucleotide polymorphic variants were obtained from 2010-8 phase 3 release of the HapMap project [109]. I considered the variants from African ancestry only, belonging to populations YRI (Yoruba), LWK (Luhya), MKK (Maasai). I also filtered for variants occurring in my intergenic non-coding sequences, resulting in a total of 1,659,929 variants.

Methods

Statistical framework to model substitution probabilities

To explain my approach for modeling nucleotide substitution probabilities observed in a given population, I will first describe a simple model that does not take into account local sequence context, then build upon this simple framework by incorporating additional features to model nucleotide substitution probabilities in a way that considers the impact of local sequence contexts of varying lengths. Suppose that we observe n_C occurrences of nucleotide C in the reference genome. A subset of these n_C sites will be polymorphic within the population of individuals. Let n_{CA} represent the number of sites where a nucleotide change C-to-A has occurred. Similarly, n_{CG} is the number of sites where a change C-to-G has occurred and n_{CT} is the number of sites where a change C-to-T has occurred. Then the probability of nucleotide substitution or polymorphism within the population genome-wide can be described at a given genomic site using a multinomial distribution:

$$\frac{n_C!}{(n_C-n_{CA}-n_{CG}-n_{CT})!n_{CA}!n_{CG}n_{CT}!}\alpha_{CA}{}^{n_{CA}}\alpha_{CG}{}^{n_{CG}}\alpha_{CT}{}^{n_{CT}}(1-\alpha_{CA}-\alpha_{CG}-\alpha_{CT})^{(n_C-n_{CA}-n_{CG}-n_{CT})} \quad (1)$$

where the probabilities of observing a substitution from C-to-A, C-to-G, and C-to-T are expressed as α_{CA}, α_{CG}, and α_{CT} respectively. After iterating over all possible substitutions (*i.e.*, A-to-C, A-to-G, A-to-T, C-to-A, C-to-G, C-to-T, T-to-A, T-to-G, T-to-C, G-to-A, G-to-C, G-to-T), I merged the reverse-complementary pairs (*e.g.*, A-to-C was merged with T-to-G, etc.) to yield 6 "substitution classes" as parameters for the simple model, which I refer to as the "1-mer model". This model can be naturally extended to consider the effects of local sequence context by replacing the count of n_x occurrences of nucleotide X with the count of occurrences of a particular nucleotide sequence context. For example, if I want to consider the local sequence context ACA, then I count the number times n_{ACA} that this 3-mer sequence occurs in the reference genome. A subset of n_{ACA} will be polymorphic at the middle position C within a given population. Thus, let $n_{ACA \rightarrow AAA}$ represent the number of sites where a nucleotide change C-to-A has occurred at the middle position, $n_{ACA \rightarrow AGA}$ represent the number of sites where a nucleotide change C-to-G has occurred at the middle position, and $n_{ACA \rightarrow ATA}$ represent the number of sites where a nucleotide change C-to-T has occurred at the middle position. All of these combinations represent a 3-mer sequence context in which the polymorphic middle position is flanked by fixed nucleotides A on both sides. After merging reverse complementary sequences, there are 16 unique sequence contexts (e.g. four possibilities (A, C, G, or T) for the single fixed nucleotide located 5' of the polymorphic site, and four possibilities for the single fixed nucleotide located 3' of the polymorphic site) per substitution class. Across all six substitution classes, there are a total of 96 parameters estimated under this "3-mer model". I analogously extend the size of the sequence context window to evaluate the "5-mer model" and the "7-mer model" by considering additional fixed nucleotides (2 and 3, respectively) on either side of the polymorphic site, thereby estimating a total of 1536 parameters for the 5-mer model and 24,576 parameters for the 7-mer model. For sake of comparison, I also considered a very simplistic null model that completely ignores sequence context and merges substitution classes into a single group, such that Equation 1 simplifies to a binomial distribution with a single estimated parameter.

Incorporating prior information into the nucleotide context models

We may have some existing "prior" beliefs regarding probabilities of nucleotide substitution that can be incorporated into my framework using Bayesian statistics. For example, rates of nucleotide substitution in the coding genome should be proportional to, but not exactly the same as, the rates that are observed in the non-coding genome. This prior information can be incorporated into my model as follows. Because the likelihood of my framework is based on a multinomial distribution, I utilize its conjugate prior, *i.e.*, the dirichlet distribution, for models that incorporate sequence context. For the null model, I can analogously utilize its conjugate prior, *i.e.*, the beta distribution. For inference in the intergenic, non-coding genome, I selected the objective version of the prior for analysis, with all concentration parameters (or shape parameters for the analogous beta prior) of the dirichlet prior as 1.

Log-likelihood ratio testing for model comparison

To evaluate how increasing the length of the context sequence affects competing models' fit to empirical data, I utilized a log-likelihood ratio testing procedure. First, the likelihood of the observed distribution of polymorphic sites given a specific sequence context model (null, 1-mer, 3-mer, 5-mer, or 7-mer) was calculated using the substitution rate parameters estimated using all of the data. I calculate the likelihood ratio test statistic as:

$$-2\ln(L[Data|Context\ S_1) + 2\ln(L[Data|Context\ S_2) \quad (2)$$

where S_1 and S_2 represent parameters estimate from two competing sequence context models. The test is chi-squared distributed, with degrees of freedom equal to the difference in the number of parameters between the two models (*e.g.*, comparing the 1-mer model versus the null model requires 5 degrees of freedom; comparing the 7-mer model versus the 3-mer model requires

24,480 degrees of freedom). Reported P-values are approximated analytically from the appropriate chi-square distribution using the R package (version 3.0.3).

Bayes Factor analysis for model comparison

I utilized the Bayes Factor approach, the Bayesian alternative to likelihood ratio testing, to contrast competing sequence context models against each other. I calculated the approximate posterior likelihood, using the Chib's method, on the overall data using the maximum *a posteriori* (MAP) estimates of the substitution probabilities for a specific sequence context model (null, 1-mer, 3-mer, 5-mer, or 7-mer) found before. I then calculate the approximate Bayes factor as:

$$\frac{Posterior\ likelihood\ under\ Model_2}{Posterior\ likelihood\ under\ Model_1} = \frac{Prob(Data|Context\ S_2) \times Prob(Context\ S_2)}{Prob(Data|Context\ S_1) \times Prob(Context\ S_1)} \quad (3)$$

where S_1 and S_2 represent parameters estimate from two competing sequence context models. Since I use flat objective priors in the noncoding region and the MAP and MLE estimates are similar, the approximate Bayes factor reduces to the ratio of likelihood estimates under the two models. I use the Jefferey's scale for interpreting the approximate Bayes Factors, where the ratio if greater than 100 is considered to be decisive evidence against the $Model_1$.

Regression modeling and feature selection

I hypothesized that, within a substitution class (described above), the probability of polymorphism could be predicted using a linear combination of features based on the nucleotides at flanking positions within the 7-mer context. For the analysis below, I considered the posterior probabilities generated using data from the African group (1KG). I considered each substitution class separately and created an additional substitution class for each of the three possible changes within a CpG context (*i.e.*, where the polymorphic 4th position nucleotide may change C-to-A, C-to-G, or C-to-T, but the 5th position in the 7-mer context sequence is fixed as nucleotide G),

resulting in nine substitution classes that are taken into regression modeling. For each substitution class, I considered the initial regression model:

$$Pr[X_1 \rightarrow X_2|S] = \alpha + \beta_1 p_1^C + \beta_2 p_1^G + \beta_3 p_1^T + \cdots + \beta_n p_7^T + \varepsilon \quad (4)$$

where the probability that a nucleotide changes from X_1 to X_2 is modeled using a position-base variable *p*, a set of bases (*e.g.*, {C, G, or T} where A is the reference base) denoted by the superscript for *p*, each position (= 1, 2, 3, 5, 6, or 7) denoted by the subscript for *p* within sequence context *S*, intercept α, and error term ε. I assigned A as the reference nucleotide at each position and encoded the single nucleotide present at each position as the combination of three thermometer variables (*e.g.*, 0,0,0 = A; 0,0,1 = C; 0,1,0 = G; 1,0,0 = T). Position 5 is fixed as G for substitution classes within a CpG context, enabling us to remove position 5 terms from those models. Similarly, models of non-CpG classes considered only C and T bases at position 5. Next, I examined non-additivity (*i.e.*, interactions) between nucleotides at sequence context positions. Rather than including all possible interaction terms, I employed feature selection (*i.e.*, model training and testing to select the most informative features) and incorporated these terms into the final model. I considered 2-way, 3-way, and 4-way interactions across positions within the 7-mer as:

$$Pr[X_1 \rightarrow X_2|S] = \alpha + \beta_1 p_1^C + \beta_2 p_1^G + \beta_3 p_1^T + \cdots + \beta_n p_7^T +$$

$$\beta_a p_i^w \times p_j^x + \cdots + \beta_b p_i^w \times p_j^x \times p_k^y + \cdots + \beta_c p_i^w \times p_j^x \times p_k^y \times p_l^z + \cdots + \varepsilon \quad (5)$$

where the probability that a nucleotide changes from X_1 to X_2 is modeled as described in Equation 4, and a set of additional terms related to interactions is also incorporated. Interaction terms are obtained from the product of thermometer variables *p* for bases *w*, *x*, *y*, or *z* (*e.g.*, {C, G, or T} where A is the reference base) at positions *i*, *j*, *k*, or *l* (= 1, 2, 3, 5, 6, or 7). The effect of the interaction is represented by terms β_a for 2-way interactions, β_b for 3-way interactions, and β_c

for 4-way interactions. I only considered interaction terms that involved nucleotides located at different positions within the sequence context (*i.e.*, *i* not equal to *j*, *j* not equal to *k*, and *k* not equal to *l*). I divided the genome into two distinct sets for feature selection, using all even-numbered chromosomes for training and all odd-numbered chromosomes for model testing. During training, I performed stepwise forward regression for each level of interaction in order of increasing complexity (*i.e.*, first 2-way, then 3-way, and finally 4-way). For each level of interaction, I further trained the model by sequentially incorporating interaction terms, one at a time, and evaluating whether each term improved the model using the ANOVA F-test. The most informative interaction term was added to the model at each step. I repeated this process until no additional features further improved the model (*i.e.*, all proposed features were $P > 0.001$ by the F-test). For higher-order (3-way and 4-way) interactions, I ensured that a proposed feature maintained the hierarchy constraint (*i.e.*, a selected 4-way term must bring with it all of its associated 3-way and 2-way terms). As a result of this constraint, when considering higher-order terms, I simultaneously considered any associated lower-order terms that had not been selected during prior lower-order training, thereby adding degrees of freedom to my F-test assessment. As my final model, I selected the trained model with the lowest mean-squared error, calculated via 8-fold cross-validation within each substitution class. I report Akaike Information Criteria and adjusted-R^2 values for the final model using the testing data set. Regression analysis was performed using R (version 3.0.3) using lm () for regression modeling, and the packages: leaps (v2.9), DAAG (1.20), lattice (v0.20-29), grid (v3.0.3), latticeExtra (v0.6-26), and RColorBrewer (v1.0.5).

Sourcing CpG methylation data

I obtained CpG methylation data for my intergenic regions of interest from a published whole genome bisulphite sequencing study performed on germline (sperm, oocyte) [110], blastocyst [110], blood [110] and brain [111] tissues. For each tissue, I divided the CpG sites into three bins:

(i) sites that were methylated in all samples, (ii) sites that were methylated in some but unmethylated in other samples and (iii) sites that were unmethylated in all samples. Very few sites fell into the second bin, so I excluded sites where methylation signal was inconsistent among the samples. I performed my analysis on the 7,059,740 intergenic CpG sites that were methylated and the 651,479 intergenic CpG sites that were unmethylated in all sperm samples. The same procedure was followed for samples from other tissues. I summarized the methylation signal across all samples for a tissue by calculating the mean intensity.

Sequence Motif Identification

I examined the top and bottom 10 sequences for each substitution class, and manually identified a total of 6 motifs that I tested in each substitution class, stratified by CpG context. This results in a total of (9 substitution classes) * (2 tails, high and low) * (6 motifs) = 108 total tests. Note that I required a nominal $P = 4.6 \times 10^{-4}$ (Bonferroni correction for multiple testing). I used Fisher's exact test to find the P-value associated with the enrichment of specific sequence motif using the fisher.test function in the R package (version 3.0.3). The contingency tables for the test were populated by considering the enrichment of sequence motifs in the top or bottom 1% of substitution probabilities for that specific class of change.

Sourcing recombination data

I obtained recombination rate map of the YRI population from the phase 1 release of the 1000 Genomes project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130507_omni_recombination_rates/YRI_omni_recombination_20130507.tar), and segregated my intergenic non-coding regions of interest into high (recombination rate >3 cM/Mb) and low recombination rate (rate < 0.05 cM/Mb) regions. As a result, I considered ~203 Mb of intergenic non-coding sequence as belonging to

high recombination rate region and ~494 Mb of intergenic non-coding sequence as belonging to low recombination rate region.

Human and primate divergence

I obtained human-chimpanzee and human-macaque chain and netted alignments from the golden path directories in the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/vsPanTro4/axtNet/, http://hgdownload.cse.ucsc.edu/goldenpath/hg19/vsRheMac3/axtNet/) and found divergence between the human-primate pair by calculating fixed differences between the aligned intergenic non-coding sequences at each 7-mer sequence context. I was able to align 1.06 Gb of intergenic non-coding sequences between human-chimpanzee and 0.88 Gb between human-macaque. For each 7-mer sequence context, I calculated the divergence as the ratio of total number of fixed differences between the human-primate pair, and the total number of sequence context occurrences in the aligned region.

Variants across the frequency spectrum

I defined the rare variants as those occurring only once or twice in the population, and low or high frequency variants as those with MAF greater than 1%. I only considered the variants present in 1000 genomes project belonging to the African ancestry and occurring in the intergenic non-coding sequences, and found 2,789,383 rare and 8,019,893 low/high frequency variants.

***De novo* mutations**

I only considered the *de novo* mutations from the high quality pedigree sequencing dataset of DECODE Genetics [16], that occurred in the accessible regions of the 1000 genomes project. This filtering was necessary because the original study did not describe the genome-wide regions that were “sequenceable”. I make an implicit assumption that atleast the accessible regions in the

1000 genomes project were sequenced in the original high quality pedigree sequencing study. I then found the observed *de* novo mutations for each motif class. The expected number of mutations occurring in each class was simulated under a normalized 1-mer sequence context model, such that the overall *de novo* mutation rate was fixed at 1.2×10^{-8} *de novo* mutations per generation and per sample.

Results

Heptanucleotide sequence context and variability in substitution probabilities

I hypothesized that local sequence context –the nucleotides that flank a polymorphic site– could explain the observed variability in nucleotide substitution probabilities. To test this hypothesis, I defined a statistical model (**Figure 2.1**) whereby the probability that a nucleotide substitution occurs at a genomic site varies based on (i) the identities of the nucleotides that flank the site and (ii) the size of the 5′-to-3′ local sequence context window. To minimize the impact of natural selection, I focused on intergenic non-coding regions of the genome. As the estimated nucleotide substitution probabilities were robust (**Table 2.1**), I developed a likelihood-ratio testing procedure to evaluate competing local sequence context models.

First, I calculated the likelihood of the observed data assuming a “1-mer” model, which allowed different substitution classes (*e.g.*, A-to-G, C-to-T, etc.) to occur at different rates but ignored effects of sequence context on substitution probabilities. I compared the 1-mer model to the trinucleotide (“3-mer”) sequence context model where single 5′ and 3′ nucleotides flanking the polymorphic middle position impact the rate of substitution. As expected, the 3-mer model significantly improved fit to the data (log likelihood ratio, LLR = 6,070,948, $P << 10^{-100}$, **Supplementary File 2.1**). Next, I evaluated if additional local nucleotides could further improve fit to the observed data. Compared to the 3-mer model or the pentanucleotide (“5-mer”, or two flanking nucleotides), I demonstrate that the larger heptanucleotide (“7-mer”, three flanking

nucleotides) fit the data better (both LLR > 494,212, P << 10^{-100}, **Supplementary File 2.2**). To further validate, I estimated substitution probabilities using 1,659,929 HapMap [109] variants found in my non-coding regions, and found that 7-mer context probabilities strongly correlated with probabilities estimated from 1KG data (**Figure 2.2, Table 2.1**), and provided the best fit to the observed polymorphisms (**Supplementary File 2.3**). My model recapitulates expected shifts in probabilities consistent with population histories [112] (**Figure 2.3**) and downward shift in the average substitution probability for the X chromosome [113] relative to autosomes (**Table 2.2**). Taken collectively, my analyses demonstrate for the first time that a 7-mer sequence context model explains the observed distribution of polymorphisms found in human populations.

To incorporate prior information, I developed a Bayesian formulation using objective conjugate priors for analysis of the non-coding genome. Consistent with my previous analysis, the 7-mer context model proved superior compared to all other models (Approximate Bayes Factor (ABF) >> 1,000, **Supplementary File 2.4**). In subsequent analyses, I utilize these posteriors for the nucleotide substitutions probabilities.

7-mer sequence context and variance explained in substitution probabilities

To quantify the variance in the posterior probabilities that a 7-mer sequence context model could explain, I considered each substitution class separately, plus CpG site contexts (9 classes total). I employed forward regression to select features from a 7-mer context window to predict substitution probabilities, and considered up to four-way interactions at positions within the window. Compared to single-base and position models without interactions, incorporating higher-order interactions substantially improved the fit to data (**Supplementary File 2.5**). Specifically, I found that my selected models in a separately held test data set explained a median of 81% of the variability (versus 30% from the 3-mer context) in probabilities across all substitution classes, covering 84% of all mutational events, and fitting well the probability of C-to-T substitution at CpGs (**Table 2.3**, **Figure 2.4**). While I identified a common set of interactions across classes

(**Supplementary File 2.6**), many common features did not always impact substitution probabilities in the same way, and others were class specific. These observations indicate that core and class-specific features based on sequence context are predictive of nucleotide substitutability.

Methylation intensity and substitution probabilities at CpG sites

The spontaneous deamination of 5-methylcytosine at CpG sites results in ~14-fold higher rates of C-to-T substitutions generally [28, 104]. Although a previous report indicated that divergence at CpG varies as a function of local context, the focus was on introns, and did not consider population-level polymorphisms in humans [114]. Thus, I hypothesized that surrounding sequence context further influences the probability of nucleotide substitution at CpGs, and examined the C-to-T substitution class within the subset contexts that contain CpG at position 4 and 5 in the 7-mer. Simulations using a model that ignored additional genomic context, or considered the 3-mer context (**Figure 2.5**), using a fixed CpG substitution probability generated significantly less variability in 7-mer CpG substitution probabilities than was empirically observed (empirical $P << 10^{-10}$, **Figure 2.4**). These data indicate that (i) not all CpG sites accrue substitutions at the same rate and (ii) that the sequence context surrounding CpG sites correlate with biological features or mechanisms that influence this rate.

To explore the possibility that the excess variability depends upon variation in methylation intensity across sequence contexts, I reanalyzed whole-genome bisulfite sequencing data obtained from germline and other tissues of healthy individuals [110, 111]. Comparing the CpG sites that are consistently methylated versus consistently unmethylated across subjects, I observed as expected that methylation correlates with an increase in the probability of C-to-T substitution ($P << 10^{-100}$, **Figure 2.6**). Unexpectedly, when I compared the methylation intensity in sperm at 7-mer CpG contexts with the probability of substitutions, I found a positive but imperfect correlation ($R^2 = 0.33$, $P < 10^{-90}$, **Figure 2.7**), with similar results in other tissues (**Figure 2.8**),

noting instances of methylation status decoupled from substitution probabilities. For example, nearly every genomic instance of the sequence contexts GTACGCA and GATCGCA showed consistent methylation signals (both methylated in >94% of occurrences in sperm), the probability of C-to-T transition was more than two-fold different for these two contexts (0.148 vs. 0.07, respectively). These data are consistent with the hypothesis that local context features beyond DNA methylation influence probabilities of C-to-T transitions at CpG sites, though I cannot exclude the possibility that sub-tissue methylation differences could explain these patterns.

7-mer context model and novel mutation promoting motifs

I next investigated the substitution probabilities for 7-mer contexts partitioned by substitution class (**Figure 2.9**, **Supplementary File 2.7**). First, I noted that several classes, C-to-A, and C-to-G in addition to C-to-T, appeared to segregate as mixtures of two distributions, explainable by CpG effects. These observations are consistent with studies demonstrating elevated substitutions at CpGs in humans [115], though this early work was not powered to measure context dependencies surrounding CpG sites as I am here. As the methylation transition state intermediate 5-formylcytosine can induce spontaneous C-to-A or C-to-G substitutions [116], one possibility is that methylation also elevates these rates in this context. I next determined if local sequence context motifs –analogous to but beyond CpG dinucleotides– correlate with variable substitution probabilities across classes. I noted that poly-CG sequences in the lower tail of C-to-T substitutions for the CpG context were enriched ($P < 10^{-16}$, **Table 2.4**). This observation is consistent with previous reports [117] as this context is found proximal to genes (**Figure 2.10**) and is associated with lower methylation intensities (**Figure 2.11**). In the upper tail of the A-to-T substitution class, I observed a poly-T+poly-A motif in the outlier sequences ($P < 10^{-5}$, **Table 2.4**). I also observed a similar quad-A motif in the lower tail of the A-to-G class ($P < 10^{-10}$). One possible mechanism that may contribute is the 'slippage' of protein machinery during DNA replication [118]. My analysis also revealed motifs without an obvious contributing mechanism.

First, in the upper tail of CpG rates, I observed enrichment of a TACG motif ($P < 10^{-10}$, **Table 2.4**) that was strongly methylated (**Figure 2.11**), but curiously, a similar motif shifted by one position was enriched in the lower tail of the A-to-C class ($P < 10^{-4}$). Second, the ApT dinucleotide was found to elevate the substitution probabilities (**Figure 2.9**) for the A-to-G ($P < 10^{-25}$) and A-to-T classes ($P < 10^{-17}$), though not statistically significantly so for A-to-C. Finally, I observed a CAAT motif also enriched in the upper tail of the A to G substitution class ($P < 10^{-53}$), reported in an earlier study of dbSNP variants [119]. These latter cases indicate potentially new mechanisms contributing to elevated nucleotide substitutability, not documented by the commonly utilized trinucleotide context model. As a final robustness analysis, keeping in mind limitations due to variant ascertainment, I estimated the substitution probabilities using HapMap variants and found similar mutation promoting motifs across substitution classes (**Table 2.5**).

Substitution probabilities and germline mutation rate

If the estimated non-coding substitution probabilities reflect properties of mutation, one would expect that these rates should (a) not influenced by rates of recombination (assuming recombination is not mutagenic) (b) strongly correlate with rates of species divergence [120], (c) be consistent for both rare and common genetic variants, and (d) also be reflected in *de novo* mutational events. I explored each of these predictions in turn. First, I estimated the 7-mer substitution rates from all intergenic non-coding variants separately for high and low recombination rate regions, and found a strong correlation between the two ($R^2 = 0.97$, $P << 10^{-100}$, **Figure 2.12**), indicating that substitution probabilities estimated from the non-coding genome are consistent across high and low rates of recombination. Next, using human-chimpanzee and human-macaque alignments over intergenic non-coding sequences, I found a strong correlation between divergence and substitution probabilities for my 7-mer contexts (both $R^2 = 0.96$, $P << 10^{-100}$, **Figure 2.13**, **Table 2.6**). I then estimated 7-mer probabilities from all intergenic non-coding rare variants (singletons and doubletons) separately from low and high frequency variants (>1%),

and found a strong correlation (R^2 = 0.98, P << 10^{-100}, **Figure 2.14)**, as well as a superior 7-mer context fit to data across variant frequencies (**Supplementary File 2.8**). Finally, I obtained 4,748 *de novo* mutational events from a high quality pedigree sequencing dataset on 78 parent-offspring trios [16]. I tested for the presence of motifs I identified in **Table 2.4** around *de novo* events, and observed a significant enrichment (**Table 2.7**). Taken collectively, these findings provide additional validation for the hypothesis that my substitution probabilities capture features of germline mutation.

CHAPTER 3: Substitution rates in the coding genome and clinical applications

Introduction

The genetic variation spectrum in protein coding regions, unlike intergenic noncoding regions is strongly affected due to forces of selection [121]. While purifying selection mostly purge deleterious variants which have an effect on fitness, other types of forces like balancing [44] or positive selection [122] also act on the coding genome. Therefore, the community has hypothesized and demonstrated that genes and variants with strong selective pressure on them, are more likely to be pathogenic and confer higher risk of disease [47]. An accurate model of this overall selective constraint, can aid in prioritization of genes and variants discovered from exome sequencing of different diseases [96], and also elucidate the differential evolutionary processes acting on the genome.

Previous approaches for variant prioritization either measure fixed differences in the genes between related species by comparative genomic approaches [123], changes to protein structure [124], deviation in frequency spectrum from neutral distribution [125] or combination of all of these [126]. While these are highly informative, they do not directly measure the overall selective constraint unique to human lineage and at the resolution of sequence context. A recent approach [31] aims to measure the selective constraint on genes as a function of local sequence context, but is limited in resolution. In the previous chapter, I demonstrate that a larger heptanucleotide sequence context explains more variability in intergenic substation probabilities, which captures features of *de novo* germline mutation. Here, I build on that framework and model substitution probabilities in the coding region as a function of sequence context and also estimate selective constraint. I show that a larger heptanucleotide sequence context model best explains patterns of polymorphisms in the coding genome, and identify previously unappreciated variability in the coding substitution rates. Next, I model a function of the selective constraint acting on the coding genome as a function of sequence context, by comparing the intergenic substitution probabilities

which are mostly shaped due to mutational forces with coding substitution probabilities which are shaped both by mutation and selective forces. I then develop several clinical utilities which can prioritize variants and genes identified from disease sequencing studies.

Data Access

Sourcing of coding sequences

I selected exonic coordinates of the longest transcript for each gene annotated in ENSEMBL Biomart (Ensembl Genes 75 and Homo sapiens genes GRCh37.p13). I only considered those transcripts where (i) the total exonic region length was a multiple of 3, and (ii) 90% or larger of it was present in the combined accessibility mask (version 20120824) filter of the 1000 Genomes project. For all genes of interest, I used phase information to map each genomic coordinate to a specific position on a codon, yielding 16,386 autosomal transcripts and 679 transcripts from the X chromosome.

To test my model in a different data set, SNP sites for ~4300 individuals of European ancestry were obtained from large-scale independent exome sequencing studies generated by the NHLBI GO Exome Sequencing Project, from the Exome Variant Server [92] (EVS, http://evs.gs.washington.edu/EVS/, downloaded on August 26th 2013).

Annotation of SNP variants in the autosomal coding genome

For 1KG data, I manually annotated the type of codon change caused by each variant, yielding 92,893 synonymous, 110,645 missense, and 1,639 nonsense variants (total n = 205,282) for the African group. I repeated the same strategy for the non-Africans, resulting 64,756 synonymous, 89,863 missense, and 1,591 nonsense variants (total n = 156,298) within the European group and 58,304 synonymous, 80,689 missense, and 1,378 nonsense variants (total n = 140,450) within the Asian group. For the EVS data (European ancestry), I also manually annotated the type of codon change, yielding a total of 226,833 synonymous, 388,149 missense, and 15,287

nonsense variants (total n = 636,122) distributed over the coding regions of interest. To obtain a representative spectrum of allele frequencies (and impact of background selection) observed from the smaller set of individuals found in the 1KG data, I considered only EVS variants with frequency greater than 0.03% resulting in a total of 169,659 variants.

Sourcing information about pathogenic variants

I used the Human Gene Mutation Database [58] (HGMD professional 2014.4) to identify pathogenic variants for my autosomal genes of interest, which supplied 60,504 variants distributed over 3,647 genes for 5,359 putative human disorders.

Methods

Extension of the substitution probability framework for coding regions

To model substitution probabilities for the coding genome, I utilized the statistical model developed for intergenic regions with the following modifications: First, I accounted for codon position-effects (*i.e.*, a given sequence context around a polymorphic site may occur at three different positions on a codon), which can lead to amino acid changes that may be subject to different levels of selective constraint. To model this phenomenon, I considered the probabilities for each of the three possible codon positions separately, resulting in a total of 73,728 (3 * 24,576) parameters for the 7-mer context model. Second, I utilized probabilities learned from the intergenic non-coding region model as my Bayesian prior for the coding model. The parameters for this prior include the baseline probabilities from the intergenic noncoding region as shape parameters for the dirichlet distribution, multiplied by an additional normalizing weighted constant, per the following:

$$\left(p_{S_1 \to S_2} * \frac{10}{1+e^{-n}}\right), \qquad \left(p_{S_1 \to S_3} * \frac{10}{1+e^{-n}}\right), \qquad \left(p_{S_1 \to S_4} * \frac{10}{1+e^{-n}}\right),$$

$$\left(\frac{10}{1+e^{-n}} - \left(p_{S_1 \to S_2} * \frac{10}{1+e^{-n}}\right) - \left(p_{S_1 \to S_3} * \frac{10}{1+e^{-n}}\right) - \left(p_{S_1 \to S_4} * \frac{10}{1+e^{-n}}\right)\right) \quad (1)$$

where *p* represents the intergenic noncoding substitution probability from sequence context *S* to each possible polymorphic change (1, 2, 3, or 4 represent each possible nucleotide base at the site), *n* is the number of occurrence of the context *S* in the coding region. This choice of shape parameter in the prior allowed for inference of coding substitution probabilities, while utilizing the intergenic substitution probabilities, and without the prior overwhelming the evidence observed in the coding region.

Scaling the model for use with a larger data sample

As the number of individuals sequenced increases, the observed number of polymorphic sites segregating within the dataset will also increase. To calibrate my model (built using the 1KG dataset) for use with the larger EVS dataset, I rescaled the substitution probabilities estimated using 1KG data to make them proportional to the EVS dataset. I used a constant scaling factor defined as:

$$\frac{Overall\ Substitution\ probability\ in\ the\ new\ dataset}{Overall\ substitution\ probability\ in\ the\ 1000\ genomes\ dataset} \quad (2)$$

on all substitution probabilities in the new dataset.

Bayes Factor analysis for model comparison in the coding region

I utilized the Bayes Factor approach, the Bayesian alternative to likelihood ratio testing, to contrast competing coding sequence context models against each other. I compared the 7-mer model with codon position effects and priors from noncoding region as described before, against the basic 3-mer model with no codon position effects and with a flat objective prior. The approximate posterior likelihood, using the Chib's method, on the overall coding data was then calculated using the maximum *a posteriori* (MAP) estimates of the substitution probabilities for the

two coding sequence context models as found before. I then calculate the approximate Bayes factor using Equation 3 from Chapter 2. For the 7-mer model the probability of parameters is found using the dirichlet distribution function in the gtools (v3.4.1) package in R (v3.0.3). Since I use flat objective priors in for the 3-mer model so the probability of parameters reduces to calculating the normalizing beta function in the dirichlet distributions. I use the Jefferey's scale for interpreting the approximate Bayes Factors, where the ratio if greater than 100 is considered to be decisive evidence against the $Model_1$.

Simulating variability in substitution probabilities within all AA classes

To simulate the distribution in variability for substitution probabilities within different amino acid substitution type, I randomly distributed the number of observed substitutions within the type using a fixed rate model. I then calculate the respective 7-mer probabilities using my multinomial distribution model for the randomization, and use those to tabulate the variance across different amino acid substitution types. 10^6 simulations are used to generate the distribution of substitution probabilities.

Measuring the effects of selection on the probability of polymorphism

To minimize the effects of selection on initial estimates of substitution probabilities, I selected intergenic non-coding intervals for model development. Assuming that the mechanisms that introduce new mutations into coding regions are similar to those at work in the non-coding genome, I inferred that the relative ratio of coding-to-non-coding substitution probabilities could indicate natural selection occurring in the coding genome. Furthermore, I expected that the rates of certain types of amino acid change should be less frequent than others (*e.g.*, on average, I expect to observe non-synonymous changes less frequently than synonymous changes) as a result of background selection. To quantify the effect of selection on substitution probabilities, I measured the $\log_{10}$ ratio of coding-to-non-coding substitution probabilities using all coding variants (n = 205,282) observed in the 1KG African group. Estimates for coding substitution

probabilities were uncertain under certain conditions, owing to a limited number of a given variant type for a particular 7-mer context. Thus, rather than use my MAP estimates for these sequences contexts, I simulated the substitution probabilities from the beta distribution using a 3-mer context model extended to the coding region. I then calculated the log-ratio of the intergenic non-coding substitution probability to the mean obtained from simulation.

Calculating tolerance scores for genes

Using my estimates for substitution probabilities in the coding genome, I performed simulations using the standard multinomial distribution for each sequence context to define the distribution of polymorphism levels expected for each gene based on my model. I then normalized the difference between the observed levels of polymorphism and those generated from my simulations, to obtain gene tolerance score defined as:

$$\frac{(\mu_{NS} - n_{NS})}{\sigma_{NS}} \quad (3)$$

where μ_{NS} and μ_{NS} represent the mean and standard deviation of nonsynonymous polymorphisms generated from simulations based on my model, and n_{NS} is the empirical number of nonsynonymous polymorphism observed in the data. A positive gene score in Equation 3 indicates that the number of observed substitutions is fewer than expected, and serves to identify genes experiencing stronger than average purifying selection. In my analysis, I determined gene scores for the African, European, and EVS populations.

Categorizing genes

I subdivided genes into various categories – *i.e.*, essential genes (where the mouse homolog knock-out is lethal), ubiquitously expressed genes, genes with known phenotypes described in OMIM, immune-related genes, keratin genes, olfactory genes and those belonging to several neuropsychiatric diseases. The dataset from [127] was used to find the first two categories, while

[125] was used to classify OMIM genes. OMIM sub-categorizes genes according to mutational models, including *de novo*, dominant, haploinsufficient, or recessive. In my analysis, I merged OMIM's *de novo*, dominant, and haploinsufficient categories, treating them as a single category. I used the DAVID ontology database [128] (version 6.7) to classify immune-related, keratin, and olfactory genes. I considered the gene list published in the latest *de novo* sequencing analysis papers of Autism [52], Epilepsy [103], Intellectual disability [129–131] and Developmental disorder [132], as the gene set belonging to these diseases. I merged the gene lists of the aforementioned diseases, treating them as single category belonging to "All Neuropsychiatric disease".

AUC comparison between competing gene scores on different gene sets

I used the receiver operating characteristic (ROC) curve to compare the performance of my gene scores against previously annotated scores [31, 125] for classifying genes into the gene sets I described above. Since, the Petrovski et. al. scores were originally released for HGNC gene ids, I was only able to convert 16,910 genes out of a total of 16,957 to corresponding ids in ENSEMBL format. Similarly, the Samocha et. al. approach only identified 1,003 genes to be intolerant and released their scores for Refseq gene ids, so I was able to map 997 genes only to corresponding ids in ENSEMBL format. Moreover, for a uniform comparison between different approaches, I only considered the previously annotated scores for autosomal genes that I identified before (i.e., which passed the stringent quality criteria of sequencing in the 1000 genomes project). I fitted a linear classifier using the three different gene scores, on each gene set and found the area under the curve (AUC) for each. The linear model was fitted using the glm function (with binomial family parameter) in R (v3.0.3). The performance of the models on different gene sets was evaluated using the pred and performance functions (with auc as a parameter) using the ROCR (v1.0-5) package.

Calculating tolerance scores for amino acids

Using my estimates for substitution probabilities in the coding genome, I performed simulations using the standard multinomial distribution for each sequence context to determine the expected number of changes for a specific amino acid within a given gene. Within a given gene, I normalized the difference between the observed numbers of amino acid changes at a specific codon versus the number of changes expected from simulation using the equation:

$$\frac{(\mu_{AA} - n_{AA})}{\sigma_{AA}} \quad (4)$$

where μ_{AA} and σ_{AA} represent the mean and standard deviation of the specific amino acid replacement polymorphisms generated from simulations based on my model, and n_{AA} is the empirical number of amino acid replacement polymorphisms observed in the data. I consider the normalized value in Equation 4 as the final tolerance score for that amino acid within the given gene. I interpret a positive amino acid (AA) tolerance score to indicate that the observed number of changes for that specific amino acid within the given gene was *even fewer* than expected. Thus, the AA tolerance score serves to identify amino acids experiencing stronger than average purifying selection. Moreover, since the AA scores measure the tolerance of a gene at an amino acid level, they further improve the resolution of the gene scores, which measure the overall tolerance in a gene. In my analysis, I determined AA tolerance scores for the African population.

Application of Gene and Amino acid scores on Autism *de novo* mutations

I used the *de novo* sequencing data for Autism spectrum disorder [90], to test the efficacy of my gene and amino acid score approach in identifying and prioritizing novel genes and variants associated with Autism. I found the *de novo* mutations belonging to cases and controls separately for each of my genic sequences of interest and further classified them into synonymous, missense, nonsense, splice and indel categories only. As a result, I considered a total of 2,171

mutations in 2,508 cases and 1,421 mutations in 1,911 controls, belonging to my genic sequences of interest.

For a uniform comparison of gene scores across different approaches [31, 125], I only considered the top 752 intolerant genes identified from each approach. I choose 752 genes because this was the number of intolerant genes identified in [31], which mapped to my autosomal genic sequences of interest (i.e., which pass the stringent criteria of sequencing quality in the 1000 genomes project). I used the Odds ratio to find the burden of *de novo* mutations in cases as opposed to controls, in the set of intolerant genes. Fisher's exact test was used to compare the significance of burden.

The amino acid scores was found on known Autism genes identified in the latest *de novo* sequencing paper [52], and compared with (a) all mutations in controls or with (b) all mutations in cases belonging to non-Autism genes. All statistical comparisons were performed using the Wilcoxon sum ranked test. Similar analysis was also performed on genes with a higher burden of functional (missense, nonsense changes for which amino acid scores are generated) *de novo* mutations in cases as opposed to controls.

Results

7-mer sequence context and variability in exonic substitution probabilities

Assuming that the processes that generate spontaneous mutations apply uniformly across the genome, I hypothesized that sequence context could explain variability in substitution probabilities in the coding genome. I therefore extended my initial framework (**Figure 2.1**) to the coding genome by (i) using information obtained from my model on the non-coding genome as prior and (ii) allowing for context dependence of codons and local sequence context in my estimates of substitution probabilities to accommodate purifying selective pressure [121]. My new model substantially improved the fit to the data compared to either 3-mer sequence context

models with or without codon context (ABF >> 1,000, **Table 3.1**). To further validate, I tested my model on a different large scale exome-sequencing dataset from ~4,300 individuals [92], and noted that my 7-mer model fit patterns of exonic polymorphisms better than competing models (ABF >> 1,000, **Table 3.1**). These results demonstrate for the first time, that a broader sequence context –beyond simple codon or trinucleotide context– captures the forces that shape variability in nucleotide substitutions in the coding genome.

I then examined the posterior distribution of substitution probabilities for all contexts stratified by the type of amino acid substitution (**Figure 3.1**, **Supplementary File 3.1**), and found excess variability in each class than expected under simulation (**Supplementary File 3.2**). Next, I enumerated the substitution probability profiles for each amino acid change, and found certain nonsense and missense substitution probabilities to be higher than synonymous levels (**Figure 3.2**), partially explained by CpG contexts. These observations caution against the practice – invoked in rare-variant association tests– of ignoring codon and sequence context when testing for the burden of functional substitutions. My results here demonstrate that functional substitutions may not be equally likely or tolerated with respect to purifying selection.

Application: 7-mer context model and power to identify pathogenic variants

I now turn to applications of my model to improve the interpretation of variation discovered by clinical re-sequencing. Efforts to prioritize variants from such studies often rely on classifying variants that are deleterious with respect to population genetic fitness, hypothesizing that such variants are more likely pathogenic [133]. As my coding substitution probabilities are influenced both by forces of mutation (estimated from the non-coding genome) and selection, I hypothesized that the ratio of these probabilities quantifies the action of selective pressure, and could be used to prioritize pathogenic variants. To test this hypothesis, I calculated the log ratio of intergenic non-coding and coding substitution probabilities, defined as sequence constraint score, for missense (n = 48,450) and nonsense (n = 12,054) variants present in the Human Gene Mutation

Database [58] (HGMD). I observed that the distribution of sequence constraint scores for HGMD variants was shifted towards larger values (intolerance) compared to 1KG variants ($P << 10^{-100}$, **Figure 3.3**), compatible with the "intolerant variant, pathogenic variant" hypothesis. Moreover, the distribution of scores based on my 7-mer model was further shifted towards intolerance with a thicker tail, compared to a 3-mer model ($P << 10^{-100}$, **Figure 3.4**). These data demonstrate that a coding model that includes codon and a 7-mer context improves identification of variants that are potentially pathogenic.

Application: Gene scores for intolerance to functional changes

Several groups have argued that the power to identify causal disease genes from clinical resequencing data could be enhanced by incorporating estimates of selective constraint on genes [31, 125, 127]. The underlying hypothesis behind this concept is that genes that are under selective constraint are more likely to have functional consequences and are therefore most likely to be pathogenic and have fewer functional variants ("intolerant gene, pathogenic gene"). The community has successfully applied this concept to neurodevelopmental and psychiatric disorders [134], however the existing approaches have not incorporated the 7-mer sequence or codon context in their models.

Therefore, I applied my 7-mer coding substitution probabilities to develop an intolerance score (**Supplementary File 3.3**) quantifying the difference between the expected and observed number of functional variants at a gene, with higher scores consistent with functional constraint. To further validate, I found gene scores on a separate, larger exome sequencing data set and observed a strong correlation between the two (**Figure 3.5**). I found that genes belonging to putatively essential or ubiquitously expressed categories, scored strongly for genic intolerance ($P << 10^{-100}$, **Figure 3.6**). In contrast, gene sets representing Keratin and Olfactory categories were found to be highly tolerant of functional changes (**Figure 3.6**). Next, I applied this to OMIM genes or known genes behind several neuropsychiatric disorders like Autism [52], Epilepsy [103],

Developmental disorder [132] and Intellectual disability [129–131], and found them to have significantly higher intolerance scores ($P << 10^{-100}$, **Figure 3.6**). I then compared my gene scores to previously reported scores (**Figure 3.7**), and found that my approach improved classification or performed comparably to other approaches [125] for genes in each set, including the disease categories (**Table 3.2**). These results demonstrate that the most accurate scoring of genic tolerance to functional substitution can be achieved by modeling 7-mer sequence and coding context.

Application: Amino acid intolerance scores and prioritization of pathogenic variants

Beyond the average rate of amino acid replacement that a gene might tolerate, genes could be further intolerant to specific types of amino-acid substitutions, signifying added localized selective constraint or importance for gene functionality. Therefore, I developed a score measuring the intolerance at amino acid replacement level in a gene (**Supplementary File 3.4**), after quantifying the difference between the expected and observed number of functional variants for a specific amino acid at a gene. Across all genes represented in HGMD with a large number of putatively pathogenic amino acid changes for a specific substitution, I found they segregate larger intolerance scores for that amino acid (empirical $P < 10^{-10}$). Moreover, a gene might score "tolerant" for functional substitution, but intolerant for specific amino acid changes. For example, Von Willebrand Factor (*VWF*), a blood glycoprotein involved in hemostasis, is tolerant to substitution overall (within top 8% of gene tolerance) but intolerant to cysteine substitution (within top 3.5% of cysteine intolerance). This data is consistent with a causal mechanism for von Willebrand disease; protein misfolding when cysteine residues are substituted [135]. Here, I note that 5,652 genes segregate a profile similar to *VWF*: average genic tolerance, but amino acid intolerance.

Application: Interpretation of *de novo* mutations discovered in Autism

Autism spectrum disorder is a disease with complex etiology, and recent efforts have aimed to identify *de novo* mutational events that may contribute to disease. To highlight the utility of gene [31, 125] and amino-acid scores, I applied them to interpret *de novo* mutations collected from 2,508 Autism spectrum disorder [90] cases and 1,911 control family trios. First, I found that the most intolerant genes based on my gene score segregated a significant burden of *de novo* mutations in cases as opposed to controls (OR = 1.66, P < 0.0004, **Figure 3.8**), even after removing known autism genes [52] (OR = 1.54, P < 0.001), and similar, though slightly attenuated burden using other scores (**Figure 3.8**). Next, I found that the average amino acid scores for *de novo* mutations at Autism genes in cases was higher (more intolerant) than that found in controls, or at other genes in cases (P = 0.002, **Figure 3.9**). I further observed higher (intolerant) average amino acid scores for variants in genes with a positive variant burden in cases, relative to controls (+2 or +3 allele count excess in cases, both P < 0.01, **Figure 3.9**). Finally, several genes from the excess allele count set stood out with amino-acid specific intolerance (all within top 4 percentile of intolerance): *MYO9B*, *WDFY3*, *NAV2*, *STIL,* and *SCUBE2.* Aside from *WDFY3*, these genes are generally 'tolerant', based on their gene-score, indicating utility of sub-gene wise measurement of functional intolerance. While *MYO9B* has been implicated in autism [52] and *WDFY3* deletions in a murine model has been shown to cause Autism like symptoms [136], my analysis points to the remaining candidates for future follow-up.

CHAPTER 4: Understanding RB pathogenesis using a mutation rate modeling algorithm

Introduction

Retinoblastoma (RB) is a cancer of the developing retina and occurs in both hereditary (~45% of cases) and sporadic (non-hereditary) forms [60]. The hereditary form of RB with the distinctive clinical features of bilateral tumors and a younger age at diagnosis is associated with the presence of germline mutations in the tumor suppressor retinoblastoma 1 (*RB1*) gene [137]. Knudson proposed a "two hit" model for the hereditary forms of RB: the first hit comprises a germline mutational event present in all cells; the second, an additional somatic genetic event within a retinal cell, overall which results in bi-allelic inactivation of *RB1* [64]. Germline mutation can either be transmitted or occur *de novo* in the affected proband with bilateral RB. Unlike transmitted variants which are present in the previous generation, *de novo* germline mutations can arise prior to conception on one of the parental genomes, or occur post-conception during early embryonic development [138].

Studies of *de novo* mutations in bilateral RB offer a new lens in which to understand the pathogenicity of specific classes of mutations. For the most part, previous studies have focused on transmitted germline mutations or mutations in bilateral probands not confirmed as *de novo*. These studies have substantially advanced our understanding of RB pathophysiology and clinical interpretation of new mutations by identifying regions or domains within *RB1* with the greatest (or least) mutational burden [67, 68, 139, 140]. However, transmitted variants themselves may not always reflect truly pathogenic mutations, owing to allele frequency differences across human populations (*i.e.*, admixture) [141]. Moreover, *de novo* mutational studies allow complete discovery of the spectrum of novel mutations that may occur in a single generation. Thus, studies of *de novo* germline mutations offer additional, orthogonal information toward understanding the mutational processes and associated pathogenic burden of new mutations that ultimately result in RB [6]. One current, open question is if *de novo* mutations occur uniformly over *RB1*, or instead

localize to specific codons, sequence contexts, or protein domains. Based on Knudson's model, we expect a higher frequency of *de novo* mutations that result in putative loss-of function (LoF) in *RB1* in patients ascertained for RB, which has been previously shown [67]. Numerous studies have reported that the preponderance of nonsense mutations occur at CpG sites in *RB1* [67, 68, 70, 140], with one study documenting a 90-fold higher rate of nonsense mutations at CpG transitions compared to non-CpG transversions in *RB1* of bilateral RB patients [70]. These observations could suggest a role of CpG sites in generating nonsense mutations [142], with deamination of hyper-methylated CpGs proposed as a mechanism to explain this "apparent" excess [139, 143, 144]. However, an alternative model could explain the observed frequency simply by disease ascertainment of loss-of-function mutations coupled with the high mutability of CpG sites due to deamination events [3]. In addition, numerous splice-site mutations have also been observed in *RB1* [68, 140, 145], many of which have been shown to result in exon skipping [145]. However, it remains to be quantified if all of these events are equally pathogenic. Finally, point mutations have been observed as frequently mutated at specific codons, which includes R661W [71, 72, 146]. This codon falls within the pocket domain in RB [147], an important domain that facilitates binding of *RB1* with downstream targets to regulate cell cycle. However, to my knowledge, enrichment of mutations at this or other codons in *RB1* has not been *statistically* quantified.

Investigating these questions for RB has been limited by the lack of a statistical model that captures mutation rate variability in human genomes at base-pair resolution. In the following, I utilize my previous work on models for nucleotide substitution using local sequence context, which explains a substantial fraction of variability in mutation rates observed in human populations [29]. I report a generalized algorithm to measure the enrichment of mutations beyond what is expected from this model, at base pair resolution. Using this approach, I performed exon- and sub-exon-level *de novo* mutational burden analysis to quantify the relative frequency of nonsense, splice-site, and missense mutations across *RB1*, in contrast to previous approaches

[67, 70, 139] that mostly binned the mutations into categories and reported simple abundance. I show that the previously reported excess of nonsense mutations in *RB1* at CpGs is compatible with the elevated rate of mutation at those sites, refuting a specific pathogenic mechanism in RB. Moreover, I report the enrichment of essential splice-site donor mutations at exon 6 and 12, but depletion at exon 5, indicative of previously unappreciated heterogeneity in relative penetrance across this type of putative LoF mutation. Finally, my approach confirms a statistically significant excess of mutations found at R661W in bilateral RB, as a hotspot for missense mutations with lower penetrance.

Data access

Patient samples

Patients included in this study were recruited as part of a research protocol between 1998 and 2011 from pediatric oncology clinics within North America. The *de novo* mutations presented here were identified from 642 children in the Genetic Diagnostic Laboratory at the University of Pennsylvania. These samples represent bilateral RB cases without family history, and where both parental DNA sample was available. Parental DNA samples were tested for the mutations identified in the respective affected child to rule out familial cases, and to unambiguously establish the presence of *de novo* mutational events. Of the 75 sporadic bilateral cases identified previously [66], only 23 samples overlap (*i.e.*, had parental samples also submitted/available). The Institutional Review Board of University of Pennsylvania approved our screening protocol.

DNA isolation and Sequencing

The isolation of DNA, PCR amplification of *RB1* sequences, and Sanger sequencing of amplified PCR products was performed as previously described [66]. Primer sequences used for amplification are available on request.

RB1 genic sequence region

I considered the genic sequence of *RB1* with accession number L11910 in the GENBANK database. Only exons 2 to 27 in *RB1* were analyzed; exon 1 was excluded to the presence of a cryptic start site [70, 148]. I also analyzed 50 base pairs on both 5' and 3' ends of the exon. Six base pairs on the 3' end of the exon were defined as donor essential splice sites, while 2 base pairs on the 5' end were defined as acceptor essential splice sites. The remaining nucleotides, from position 7 to 50 on the 3' end of the exon, and from 3 to 50 on the 5' end of the exon were defined as intronic sites. As a result, I analyzed a total of 5,460 nucleotide bases in the gene, out of which 2,787 were from protein coding region, 2,457 intronic and remaining 216 belonging to essential splice sites.

RB mutational data

A total of 571 mutations were identified, which included 289 point mutations. Furthermore, I considered missense, nonsense, essential splice, and intronic mutations that falling in the *RB1* sequence region defined per the above, that passed quality control. As a result, 268 mutations falling in my region of interest were analyzed.

ExAC variants

I only considered the singleton variants (missense, nonsense, essential splice and intronic) in the non-Finnish European populations from the ExAC dataset. I initially downloaded all variants in "ENSG00000139687" gene id from ExAC, including only mutations that were observed once (singletons). As a result, I analyzed 149 singleton variants falling in my region of interest as described above.

RB mutational data from collaborators

I independently received nonsense mutational data from a recent publication of germline *de novo* mutations in RB [70]. I analyzed 100 variants from this dataset that were present in my region of interest as described above.

LOVD variants

I queried the variants present in the 2015 release of the Leiden Open Variation Database (LOVD: http://rb1-lsdb.d-lohmann.de/variants.php?action=search_unique&select_db=RB1) for in the *RB1* gene. I only reported the results from the point mutations present in the database.

Methods

Analysis of the total number of mutations discovered

Unlike the noncoding region, previous studies [16] have reported a higher *de novo* mutation rate of ~1.5×10^{-8} mutations per base pair per generation in the coding region. Since I consider a total genomic territory of 5,460 nucleotide base pairs in RB1 gene and sequenced 642 individuals (or 1,284 haploid chromosomes), I expect a total of 0.1 *de novo* mutations in our sample. This is calculated by multiplying the *de novo* mutation rate (1.5×10^{-8}) with the total genomic territory (5460 base pairs) and total number of haploid chromosomes sequenced (642 x 2). Since, I observed 268 non-silent *de novo* mutations and I expect 0.1 *de novo* mutations, I report extreme statistical significance after simulations from a Poisson distribution with fixed parameter as the expected mean of 0.1.

Analysis conditional on a set of observed mutations

The majority of analyses presented in the paper focused on generating the expected number (and variance in) mutation number, conditioned on a specific type of event or sub-sequence with *RB1* where a set of events had occurred. In the case of mutations identified in RB probands, this involved distributing all (or a subset of) n=268 *de novo* mutations we discovered by re-sequencing *RB1*. For the comparison to ExAC, this involves distributing all (or a subset of) n=149 singleton variants I identified in non-Finnish Europeans, as an admittedly imperfect proxy for *de novo* events. My procedure involved three steps:

Step One: Select genomic territory and observed mutations that fall in region of interest. For results where n=268 mutations were distributed, I considered all of the available genomic territory that was re-sequenced and filtered from my discovery effort (*i.e.*, a total of 5,460 bases, as described above). Here, if the base pair position did not result in a desired type of mutation, that base is excluded. The number (and type) of mutations that are subsequently distributed was based on those actually discovered within the specified territory. Finally, I assumed *de novo* mutations located in any/all positions in my territory was always able to be discovered, if present.

Step Two: Distribute mutations on sequence according to context model. The probability of mutation at each base pair of the genomic territory selected in Step One is provided by my 7-mer sequence context based substitution probabilities, which were estimated from the non-coding genome in my prior work [29]. Briefly, a nucleotide base can change into one of three other bases (*e.g.*, nucleotide C can change to A, G, or T) with different substitution probabilities based on the type of change. Depending on the codon and position context, this nucleotide change can result in one of many types of coding changes (*e.g.*, nonsense, splice-site, etc.). The type of mutations selected in Step One determines which of these three nucleotide base changes at the position is considered. For example, if only nonsense mutations were selected in Step One, I would consider only the base pair positions and subset of possible nucleotide changes in each base pair that result in a nonsense mutation. Once all probabilities across base pairs have been identified, I then normalize by the sum of all probabilities so that the total at all eligible bases where a change could occur in the gene is 1. For a given simulation and the total number of mutations selected in Step One, each is distributed across the gene from a multinomial distribution with probabilities as estimated before.

Step Three: Determine Empirical Significance. For each comparison, I performed 1,000,000 simulations to determine the empirical distribution of mutation count found at the type or sub-sequences of mutations specified in Step Two. Empirical p-values for significant enrichment

(deficiency) were determined by counting the number of times that the simulations had a value greater (less) than or equal to the observed number of mutations in that class.

Rates of different classes of mutation, relative to nonsense mutations

I first calculate the ratio of observed to expected mutations in a category after distributing all 268 *de novo* mutations at all eligible bases and possible changes (any change except those resulting in a synonymous mutation) using my algorithm described before. Next, I normalized this ratio by dividing it with the mean for nonsense category. This results in setting the mean of observed to expected variants for nonsense category as 1. I then plot the mean and standard error of this ratio for each category of mutations. The different distributions of this observed to expected rate are compared using a standard 2 sample t-test.

Results

Re-sequencing of sporadic bilateral RB patients

To quantify the role of *de novo* mutations in the pathophysiology in RB, we re-sequenced *RB1* in 642 cases presenting sporadic (*i.e.*, without family history), bilateral RB and their parents. The targeted resequencing included all exons of *RB1* as well as 50 base pairs of intronic sequences on either side of exons. For statistical modeling purposes, I focused on single base point mutations and excluded individual who carry a frame-shift or in-frame insertion-deletion mutations. After variant calling followed by quality control, I identified 276 *de novo* germline, single base point mutations. Owing to an alternative start codon in exon 1 [70, 148], my subsequent analyses focus on the remaining exons, resulting in 177 amino-acid altering mutations, 86 in essential splice-sites, and 5 mutations found in introns outside of essential splice-sites (total of 268 *de novo* events, **Supplementary File 4.1**). Consistent with the causal role of *RB1*, the discovery of 268 *de novo* mutations in 642 RB probands is highly unusual (Expected number of variants = 0.1, $P << 10^{-10}$). Furthermore, I observed more nonsense and

essential splice-site mutations than missense or intronic mutations, expected given the pathogenic nature of loss-of-function (LoF) mutations in *RB1* (**Table 4.1**). For a population-level comparison, I contrasted my mutational profile to the data obtained from the Exome Aggregation Consortium (ExAC) [149], consisting 60,706 individuals re-sequenced for the exome. Here, I note that ExAC excluded childhood diseases from their aggregation, which may have excluded RB patients. As a result, I do not expect this sample to represent a completely random population sampling of mutations in *RB1*. From ExAC, I focused on singletons observed in non-Finnish populations of European ancestry (n=149 variants in >33,000 subjects, **Supplementary File 4.2**). Consistent with samples from ExAC as population-level controls with potential ascertainment against RB disease, I observed fewer loss-of function and more missense and intronic variants compared to the *de novo* mutations identified in RB probands (**Table 4.1**).

An algorithm to quantify the enrichment of *de novo* mutations

My central objective is to determine if the frequency, type, and location of *de novo* mutations in *RB1* were consistent with the number of events predicted from my local, nucleotide sequence context model for mutation rate variability. For example, I expect more nonsense mutations in RB patients than my background model predicts, because (i) I ascertained individuals with RB, (ii) nonsense mutations are likely LoF, and (iii) LoF at *RB1* causes RB. To achieve this objective, I require an accurate model that captures variability in the frequency of *de novo* mutational events across *RB1* and an engine to distribute mutations in *RB1* according to this model. With these in place, I can empirically assess significance of enrichment of *de novo* mutations in exons or sub-sequences of *RB1* relative to my model prediction.

In my previous work [29] I demonstrated that an expanded sequence context model which considers three flanking nucleotides on either side of a base (*i.e.*, heptanucleotide), explains variation in germline mutation rate better than competing models of sequence context, and up to 93% of the variability in substitution probabilities. Using the sequence context based substitution

probabilities, I developed an algorithm to distribute mutations across the gene in order to generate an expected count of mutations (with variance) at all positions in *RB1* (**Figure 4.1**). With these distributions in hand, I can estimate the empirical significance conditioned on the observed number of any type of substitution in any sub-sequence(s) within the gene. As an imperfect control, I use singletons from ExAC (allele frequency of ~1/66,000, ~0.00152%) in which to compare our *de novo* events, with the assumption that these events are the youngest and have not experienced the full force of purifying selection; *i.e.*, are the closest proxy to *de novo* events segregating in (non-Finnish) European populations. In what follows, I apply my approach to study (i) the overall frequency of nonsense, essential splice-site, and missense mutations in *RB1* and ExAC, and (ii) their spatial occurrence by exon or by sub-sequence (CpG sites, domains, or codons).

Abundance of nonsense mutation at CpG sites explained by elevated mutation rate

I first investigated if nonsense mutations were distributed proportionally to the predicted rate of mutation, or alternatively localize to specific sequences, like CpGs. As a positive control, I first distributed the 268 identified mutations ascertained in RB probands and determined how many nonsense mutations were predicted from my sequence context mutational model. I found an enrichment of nonsense mutations beyond that expected from my model ($P << 10^{-6}$, **Figure 4.2**). This observation is consistent with extensive literature showing that LoF mutations at *RB1* cause RB. As a negative control, I distributed variants identified from the ExAC database, and observed fewer nonsense mutations than expected based from my model ($P = 0.0103$, **Figure 4.2**). This is also expected, as I anticipate few (if any) nonsense mutations in *RB1* observed in the general population or in ExAC participants which may have excluded RB patients.

I next examined if the subset of 150 nonsense mutations I observed were unusually distributed across exons in *RB1*. I found that, across virtually all exons, nonsense mutations occurred as frequently as my model predicts, broadly consistent with the concept that nonsense mutations

found across *RB1* are similarly pathogenic (**Figure 4.3**). The single exception was exon 27, which segregated fewer mutations than my model predicted ($P << 10^{-6}$, **Figure 4.2**). This observation is compatible with the hypothesis that nonsense mutations in exon 27 are not fully penetrant, perhaps due to incomplete nonsense mediated decay [150] or that this exon may not be integral to the etiology of RB. Previous studies have observed fewer mutations at later exons in the *RB1* gene [67], though they were unable to quantify the reduction and assess statistical significance as I am able to do here. While I observed fewer mutations at exons 25 and 26, these numbers are still compatible with my background mutational model, given the number of mutations that were discovered in re-sequencing.

Next, I examined if the subset of 150 nonsense mutations I observed were unusually distributed in amino acid type or codon contexts across *RB1*. I found that the distribution of *de novo* events by amino acid and codon context was not especially different from what my mutational model predicted (**Table 4.2**). Specifically, my model predicted a large number of C-to-T transitions resulting in Arginine to Stop mutations at the CGA codons (93 observed, 99% CI: 73 – 104, P = 0.24), presumably due to the higher mutational frequency at the CpG context [3, 30]. This analysis indicates that observed profile of nonsense mutations can be explained simply by the background rate of mutation, without a need to invoke a RB-specific mutation-promoting or pathogenic mechanism at CpG sites.

To replicate these observations, I repeated my analysis on an independent set of 100 nonsense *de novo* germline mutations in *RB1* of bilateral RB patients (**Supplementary File 4.3**). These results recapitulated the observed deficiency of nonsense events in exon 27, and my model also matched the number of nonsense mutations at CpG sites or at CGA codons relative to other nonsense sites (**Tables 4.3** and **4.4**).

Excess splice-site donor mutations in introns 6 and 12, but depleted in intron 5

I next investigated if essential splice-site and intronic mutations were distributed proportionally to the rate of substitution predicted by my context model. As a positive control, I distributed the 268 mutations ascertained in RB probands and determined how many essential splice-site and intronic mutations were expected from my sequence context mutational model. I found more *de novo* essential splice sites mutations in RB patients than predicted ($P << 10^{-6}$, **Figure 4.3**). This observation is consistent with the idea that essential splice-site mutations that are LoF at *RB1* cause RB. As a negative control, I distributed variants identified from the ExAC database and observed fewer essential splice variants there ($P = 0.014$, **Figure 4.3**). This is not unexpected: analogous to nonsense mutations described above, I anticipate few essential splice-site mutations in the general population and/or ascertainment against RB patients in ExAC participants. In intronic sequences that are found outside of essential splice sites, I observed substantially fewer events in RB patients than my model predicted ($P << 10^{-6}$, **Figure 4.3**). In contrast, I found more intronic events in ExAC than my model would predict ($P << 10^{-6}$, **Figure 4.3**). Taken collectively, these two observation suggest weak (if any) overall pathogenicity for intronic mutations outside of essential splice-sites.

I then examined if the 86 essential splice-site mutations I ascertained in RB probands were unusually distributed across introns in *RB1*. First, I found that essential splice-site acceptor mutations were not unusually distributed (**Figure 4.4**), so I focused on the remaining 63 essential splice-site donor mutations. Next, I observed no mutations in the donor site of intron 5, which was outside my model prediction ($P << 10^{-6}$, **Figure 4.5**). However, this observation is readily explainable: if we assume that essential splice-site donor mutations here result in exon skipping as seen for other splice-site mutations [145], it turns out that skipping exon 5 retains the coding reading frame, albeit with a 13 amino acid deletion (**Figure 4.6**). Therefore, this type of mutation may not result in full LoF of the *RB1* protein product, and thus, may be weakly penetrant, if at all. Next, I found that essential donor splice-site mutations in intron 6 and 12 segregated more

mutations than my model predicted ($P << 10^{-6}$, **Figure 4.5**). Previous studies [68, 140] have observed that exon 6 and 12 mutations are recurrently mutated in *RB1*, though they were unable to quantify the enrichment and assess statistical significance as I are able to here.

It is not immediately apparent why these *specific* splice-site mutations are enriched in RB ascertained patients compared to other splice donor mutations. Essential donor splice-site mutations at intron 6 and 12 result in exon skipping [145], out-of frame shift mutation, and putative LoF (**Figure 4.6**). However, essential donor splice-site mutations at other introns (except intron 5) also result in frame-shift mutations in *RB1* if exons are skipped. To further validate the observation of specific enrichment at these exons, I utilized the Leiden Open Variation (LOVD) Database [151] (Methods), a curated catalog of mutations found in *RB1*. Because variants are reported from multiple studies, where the gene territory re-sequenced and total number of individuals ascertained is not completely documented, I am limited in my ability to statistically quantify variant enrichment in LOVD as I can for our RB data. I found recurrent mutations with multiple reported variants (or fewer for exon 5) even in the LOVD database of all reported variants in *RB1* gene of patients with RB (**Table 4.5**). Moreover, the donor sequences of inton 6 and 12 also are similar to other canonical splice sequences found at other (not enriched) exons. Taken collectively, these data suggest some additional pathogenic burden of these mutations relative to other essential splice-sites in *RB1*.

Localized enrichment of missense mutations to R661W in *RB1*

I investigated if missense mutations were distributed proportionally to the rate of substitution predicted by my context model. I distributed the observed 268 mutations across the gene, and found significantly fewer missense mutations than expected ($P << 10^{-6}$, **Figure 4.7**). This observation is consistent with the model that missense mutations as a class *generally* are less penetrant for RB, contrasting against the substantially higher penetrance of LoF nonsense or essential splice mutations. In contrast, ExAC participants were not unusual in the distribution of

missense variants observed relative to my model prediction (P = 0.041, **Figure 4.7**). Taken collectively, these data suggest that, as a class, missense mutation in *RB1* are less frequently pathogenic than nonsense variants and result in fewer mutations ascertained in RB probands.

The idea that missense mutations *generally* are less penetrant for *RB1* still leaves open the possibility of heterogeneity in pathogenicity among sub-sequences of *RB1*. For example, R661W is a frequently observed mutation found in families that segregate lower penetrance [71, 72, 146]. Computational prediction tools like Polyphen2 [124] or evolutionary conservation based metrics [152] are frequently used to rank missense variants categories of deleteriousness as a proxy for pathogenicity. I applied Polyphen2 to classify all missense mutations we identified, and found most of them to be damaging (**Table 4.6**).

To further improve the resolution of these predictions, I applied my approach to identify a smaller, statistically credible subset of missense mutations implicated in RB pathogenicity. To achieve this, I distributed all 27 missense mutations I ascertained in RB probands across *RB1* to determine if these rates were proportional to my predicted mutational model. I observed a significant enrichment of missense mutations in exon 20, mapping to the known pocket domain in *RB1* (**Figure 4.8**, 8 mutations out of 27, $P << 10^{-6}$). Although the pocket domain in *RB1* gene encompasses other exons [71, 147] (*i.e.*, Pocket Domain Box A: Exons 13-17, Pocket Domain Box B: Exons 18-22), I did not observe a specific enrichment of missense mutations there (all $P > 0.01$, **Figure 4.8**). I next distributed the missense mutations within the pocket domain territory in *RB1* (n=18 missense mutations in 307 codons across the entire pocket domain). I observed an excess of missense mutation burden within exon 20 in Pocket Domain Box B near codon 661 than predicted by my model ($P << 10^{-6}$, **Figure 4.9**).

I next sought to localize the signal of the missense mutational burden within exon 20. I distributed all missense mutations I observed within exon 20 (n=8 in total), and observed an enrichment of missense mutations from CGG to TGG coding for a change from Arginine to Tryptophan (**Table**

4.7). Specifically, I found the previously observed recurrent mutation R661W (n=5 times in my sample) occurred more frequently that my model predicted ($P << 10^{-6}$). Here, I note the limited resolution of Polyphen2, as it also predicts other sites nearby as damaging (**Table 4.6**).

To place this observation in context of other missense mutations documented in *RB1*, I evaluated the frequency of n=130 missense mutations in exon 2 to 27, curated by the LOVD repository. There, the most frequently cataloged missense mutation was R661W (n=33 of 127), with the next most frequently listed as C712R (n=8 of 127), G137D (n=6 of 127), and T307I (n=5 of 13). However, when reflected against ExAC, R661W was observed only once (<0.001%) and C712R was not observed at all, consistent with putative pathogenicity of both variants. In contrast, G137D and T307I were far more frequent in ExAC (0.04% and 0.3%, respectively), suggestive of very low RB penetrance for these events. While the LOVD ascertainment is certainly complex and precludes me from formally evaluating statistical significance, these data are consistent with the importance of R661W as pathogenic and a frequently mutated position.

Relative rates of different classes of mutations found in *RB1*

Finally, I sought to quantify – relative to nonsense mutations – the rates of various sub-types of *de* novo mutations I observed in *RB1*. Assuming the penetrance of nonsense mutation is nearly full, the idea here is that if a subtype of *de* novo mutation was as penetrant as nonsense mutations, I would expect to have ascertained that subtype as frequently as nonsense mutations, proportional to the mutability of the subtype. I found that the rate of ascertainment of essential splice-site mutations was statistically lower than nonsense mutations ($P << 10^{-10}$, **Figure 4.10**), consistent with the lower penetrance of essential splice mutations due to some less pathogenic changes observed at the essential splice positions (*e.g.*, intron 5). Similarly, the rate of intronic and missense mutations relative to nonsense was substantially smaller ($P << 10^{-10}$, **Figure 4.10**). Finally, while the rates of missense mutations found in both Pocket Domain Box A and B were less frequent relative to nonsense mutations, I noted that mutations localized to Box B were more

frequent compared to missense mutations overall or in Box A (both $P << 10^{-10}$, **Figure 4.10**). Together, these data suggest a mixture of penetrant missense mutations found across *RB1*, elevated in penetrance for Box A mutations, and further elevated in Box B, the Box that also contains codon 661.

CHAPTER 5: A framework for interpreting *de novo* mutations in human disease

Introduction

Mutation creates genetic variation between individuals, and is critical for understanding population history [15], estimating evolutionary distances between species[9], detecting natural selection [153] and for discovering causal genes and variants[31, 125, 154] behind genetic diseases. Previous studies [3, 30], including ours [29] have reported significant variability in the mutation rate across the genome, and at the level of local sequence context around the polymorphic site. A striking example of the local sequence context in influencing the rate of mutation, is the spontaneous deamination of 5-methylcytosine at methylated CpG sites which causes a ~15-fold elevated C-to-T mutation rate relative to the genome [28]. Recently the community has tried to understand and measure the role of *de novo* mutation burden in several childhood disorders like Autism [52, 89], Epilepsy [103] and Congenital heart disease [53] to identify causal genes. All existing approaches [155, 156] for measuring this mutation burden at the genic level require an accurate estimate of mutation rate which captures the variability, and a systematic framework to model, measure and test for the burden of *de novo* mutations in affected probands.

Previous methods to estimate variability in the germline mutation rate can be broadly categorized into the phylogenetic [157, 158] or the parent-offspring sequencing [16] approach. The phylogenetic approach considers the neutral mutations (mostly silent mutations at few pseudo-genes) between humans and other hominid species and then under the molecular clock assumption find the coarse estimates of mutation rate. Family trio sequencing approaches, which measure the *de novo* mutations in each generation of a family, have higher resolution and are more accurate than phylogenetic estimates, but are still severely underpowered owing to sparsity of de novo mutations identified in each generation (only ~70 per parent-offspring trio). Recently, several studies have also estimated *de novo* mutation rate from millions of SNPs in the intergenic

noncoding region, which are mostly shaped due to forces of mutation [31, 89], to accurately capture the variability in the mutation rate. However, they use a less informative sequence context model for mutation rate estimates, and also a simplistic framework to model and test for burden of *de novo* mutations in affected cases. Here I address these challenges by developing a systematic framework to estimate *de novo* mutation rate at different competing sequence context models from millions of SNPs from 1KG dataset [13]. I demonstrate that a larger heptanucleotide sequence context based *de novo* rate estimates best explain variability in mutational data. I also report a higher mutation rate estimate for the coding genome and finally develop a systematic framework that allows the user to simulate or distribute *de novo* variants over genomic regions.

Data Access

De novo mutations

I only considered the *de novo* mutations from the high quality pedigree sequencing dataset of DECODE Genetics [16], that occurred in my defined whole genome accessible territory. This filtering was necessary because the original study did not describe the genome-wide regions that were “sequenceable”. I make an implicit assumption that at least the accessible regions in the 1000 genomes project were sequenced in the original high quality pedigree sequencing study. As a result, I finally considered 4,748 *de novo* germline variants over 2.53 GB of autosomal genomic territory from sequencing of 156 haploid chromosomes (or 78 individuals) for my analysis.

Whole genome territory

We assumed that the accessible regions (combined accessibility mask (version 20120824) of the 1000 genomes project) from the 1000 genomes project were “sequenceable” and of high quality. As a result, we considered 2.53 GB of autosomal genomic territory for analysis.

Intergenic non-coding region

Intergenic sequences were defined as the full set of genomic sequences that are not annotated in ENSEMBL Biomart [107] (Ensembl Genes 75 and Homo sapiens genes GRCh37.p13) and RefSeq Genes [108]. I further removed genomic regions 10 KB upstream and downstream of all such coding annotations. This is to further ensure that regulatory regions enriched near the genic annotations, and are subjected to forces of selection [46] are not considered in our analysis. Finally, I intersected this set of regions with the accessibility mask filter of (combined accessibility mask, version 20120824) of the 1000 genomes project to find the high quality set of "sequenceable" intergenic non-coding regions. As a result, I considered 0.83 GB of autosomal intergenic noncoding genomic territory for our analysis.

EVS variants

To test my *de novo* sequence context rates in the coding region, I only considered variants from the EVS data which occurred as singletons or doubletons, hypothesizing that these variants may have not been subjected to full forces of selection and represent the closest proxy to *de novo* mutations in the coding genome.

Coding transcripts

I selected exonic coordinates of the longest transcript for each gene annotated in ENSEMBL Biomart (Ensembl Genes 75 and Homo sapiens genes GRCh37.p13) and RefSeq Genes database. I considered transcripts where the total exonic region length was a multiple of 3. For all genes of interest, I used phase information to map each genomic coordinate to a specific position on a codon, yielding 16,342 autosomal transcripts and 711 transcripts from the X chromosome.

Methods

De novo rates estimation

As previously described in my work in Chapter 1 and in a recent publication [29], I first found the intergenic substitution probabilities from the 1000 genomes Phase 1 [13] data over our defined intergenic noncoding regions of interest. Within these intergenic regions, I found 7,504,983 single nucleotide polymorphic variants in the African populations and 4,878,890 variants in the European populations. Since, intergenic substitution probabilities are (a) mostly shaped due to forces of mutation and (b) I have further removed regions upstream and downstream of genes compared to our previous analysis where I showed that substitution probabilities capture properties of *de novo* germline mutation, so I assume here that these latest intergenic substitution probabilities also capture variability in mutation rate.

Next, I developed an approach to estimate *de novo* mutation rate at each sequence context from the substitution probabilities inferred before. Since substitution probabilities are a constant scalar multiple of *de novo* mutation rate, we can estimate this scalar multiple by fixing the overall *de novo* mutation rate to $1.2*10^{-8}$ mutations per base pair per generation. In Equation 1, I present our approach to find this scalar multiple.

$$\frac{Scalar\_Multiple \times \sum_{i=1}^{all\ contexts} Substitution\_Probability_{Context_i} \times Count\_Context_i}{\sum_{i=1}^{all\ contexts} Count\_Context_i} = 1.2 \times 10^{-8} \quad (1)$$

We can solve for the scalar multiple by plugging in the substitution probabilities and the corresponding counts at each sequence context. The overall idea behind this is that ratio of genome-wide expected mutations to genome-wide territory is the overall genome-wide *de novo* rate of mutation. We can find the overall genome-wide expected mutations by multiplying the mutation rate at a context (scalar multiple of substitution probabilities) to its number of occurrences. Similarly, we can find the overall genomic territory by summing the counts of all sequence contexts. Here, I note that my approach is generalizable and I can estimate the *de*

novo mutation rate at each sequence context with any other overall estimate of mutation rate. Our substitution probabilities, capture the relationship between mutation rate at different sequence contexts, and can be normalized with any overall estimate to find sequence context specific mutation rate estimates.

Comparison of competing sequence context models

Whole Genome comparison: To evaluate how increasing the length of the context sequence affects competing models' fit to the test data, I utilized a log-likelihood comparison procedure. We trained and estimated different sequence context based *de novo* germline mutation rates from the 1KG dataset and tested their fit on a separate *de novo* germline mutational dataset [16], of 4,748 mutations from sequencing of 78 family trios. The likelihood of the observed distribution of *de novo* mutations given a specific sequence context model (null, 1-mer, 3-mer, 5-mer, or 7-mer) was calculated using the statistical framework, I described before [29] in Chapter 1. I will first describe here a simple model that does not take into account local sequence context, then build upon this simple model by incorporating additional sequence context features. Suppose that we observe n_C occurrences of nucleotide C in the reference genome. A subset of these n_C sites will have a *de novo* germline mutation in our sample. Let n_{CA} represent the number of sites where a mutation C-to-A has occurred. Similarly, n_{CG} is the number of sites where a mutation C-to-G has occurred and n_{CT} is the number of sites where a mutation C-to-T has occurred. Then the probability of *de novo* mutation within our sample after sequencing *n* haploid chromosomes can be described at a given genomic site using a multinomial distribution:

$$\frac{n*n_C!}{(n_C-n_{CA}-n_{CG}-n_{CT})!n_{CA}!n_{CG}n_{CT}!}\alpha_{CA}{}^{n_{CA}}\alpha_{CG}{}^{n_{CG}}\alpha_{CT}{}^{n_{CT}}(1-\alpha_{CA}-\alpha_{CG}-\alpha_{CT})^{(n_C-n_{CA}-n_{CG}-n_{CT})} \quad (2)$$

where the probabilities of observing a mutation from C-to-A, C-to-G, and C-to-T are expressed as α_{CA}, α_{CG}, and α_{CT} respectively. This model can be naturally extended to consider the effects of local sequence context by replacing the count of n_x occurrences of nucleotide *X* with the count of

occurrences of a particular nucleotide sequence context. For example, if I want to consider the local sequence context ACA, then I count the number times n_{ACA} that this 3-mer sequence occurs in the reference genome. A subset of n_{ACA} will have mutation at the middle position C within the sample. Thus, let $n_{ACA \rightarrow AAA}$ represent the number of sites where a mutation C-to-A has occurred at the middle position, $n_{ACA \rightarrow AGA}$ represent the number of sites where a mutation C-to-G has occurred at the middle position, and $n_{ACA \rightarrow ATA}$ represent the number of sites where a mutation C-to-T has occurred at the middle position. All of these combinations represent a 3-mer sequence context in which the middle position is flanked by fixed nucleotides A on both sides. I analogously extend the size of the sequence context window to evaluate the "5-mer model" and the "7-mer model" by considering additional fixed nucleotides (2 and 3, respectively) on either side of the polymorphic site. Using this framework, I calculate the likelihood of different sequence context models on the test data of *de novo* germline mutations.

Coding Genome comparison: I further tested the *de novo* germline estimates of the exonic region on my defined coding transcripts. However, this test was performed on our defined EVS variants occurring in the coding transcripts. I first scale the *de novo* germline mutation estimates by a constant multiple, specific for each sequence context model, such that the overall number of expected variants is the same as total number of EVS variants (n = 702,935) observed in our sample. The total number of expected variants is calculated using Equation 3,

$$\sum_{i=1}^{all\ contexts} De\ novo\ mutation\ rate_{Context_i} \times Count\_Context_i \quad (3)$$

with counts calculated on my defined coding territory. Then, I use the likelihood comparison procedure defined above, to find the likelihood of competing sequence models on the EVS data. The rates are all scaled by the constant multiple, counts are calculated on my defined coding region and counts of mutations are the EVS variants observed in our sample.

De novo mutation rate in coding region

I find the *de novo* mutation rate estimate in the coding region by finding the ratio of total expected mutations in my defined coding region to the total nucleotide count there. The total expected mutations are calculated as defined above. I multiply the *de novo* mutation rate estimate at each context to its count in the defined coding region to get the total expected mutations.

I also find the *de novo* mutation rate specific to the type of change (synonymous, missense, splice, nonsense, specific amino acid) in the coding region by incorporating codon structure in my model. First, I classify each sequence context change into the desired category of substitution (i.e. it results in a change like synonymous mutation or at a specific amino acid), and then I multiply its sequence context mutation rate with its total occurrence in the category. I then sum this multiple for all eligible sequence contexts and divide it by the total count of nucleotides to find the *de novo* mutation rate specific to a substitution class.

Simulation of *de novo* mutation

In my toolkit SimDenovo, I simulate *de novo* germline mutations over any genomic region queried by the end user, and output the total and individual category specific expected mutations over that region. The user enters the number of haploid chromosomes *n* over which *k* simulations need to be performed. The central idea is to find the total expected mutations over any genomic region, after sequencing *n* haploid chromosomes.

For any such genomic region, I first find the total count of each heptanucleotide sequence context in the genomic region, and also multiply the mutation rate at each heptanucleotide sequence context with the *n* haploid chromosomes entered by the user. Next, I use the previously defined multinomial statistical framework to simulate mutations at any sequence context over *k* simulations. For each simulation, I find the number of mutations from a heptanucleotide sequence context to another. Finally, I assign the simulated mutation (for all *k* simulations) at a sequence

context, to any randomly chosen occurrence of that sequence context in the genomic region, and classify it into a specific category (missense, nonsense, synonymous, amino acid or CpG etc.). The final output to the user includes summary statistics of total and individual category specific simulated *de novo* germline mutations.

Distributing of *de novo* mutation

In my toolkit SimDenovo, I distribute *n de novo* germline mutations over any genomic region queried by the end user, and output the total and individual category specific expected mutations over that region, based on the background rate of mutation. The user enters the total number of *de novo* mutations *n* that need to be distributed over the genomic territory after performing *k* simulations. The central idea is to find expected mutations in each sub-category in the genomic region conditional on observing a total of *n* mutations.

My procedure for distributing mutations first involves finding the probability of mutation at each base pair of the genomic territory. A nucleotide base can change into one of three other bases (*e.g.*, nucleotide C can change to A, G, or T) with different mutation probabilities based on the type of change. Once all probabilities across base pairs have been identified, I then normalize by the sum of all probabilities so that the total at all bases where a change could occur in the genomic territory is 1. For a given simulation and the total *n* number of mutations, each is distributed across the genomic territory from a multinomial distribution with probabilities as estimated before. Finally, I classify the mutation into a specific category (missense, nonsense, synonymous, amino acid or CpG etc.). The final output to the user includes summary statistics of individual category specific distributed *de novo* germline mutations.

Results

Accurate and Informative estimates of *de novo* mutation rate

I hypothesized that a sequence context based approach using millions of SNPs from population level sequencing data (1000 Genomes project [13]) could be used to infer the variability in *de novo* mutation rates. To test this hypothesis, I built on my previous work [29] and defined a statistical model where the substitution probabilities were inferred for windows of different sequence context length, from population level SNPs over intergenic non-coding regions far away from genes to minimize the impact of selection. I then normalized the substitution probabilities to get mutation rate estimates at each sequence context, such that the genome-wide *de novo* mutation rate is fixed at 1.2×10^{-8} mutation per site per generation [16]. Since the *de novo* germline mutation rate is similar across humans [3, 8], I further hypothesized that the mutation rates inferred from polymorphism data across different populations result in similar estimates. In order to test this hypothesis, I robustly estimated the *de novo* mutation rates at hepta-nucleotide ("7-mer", three flanking nucleotides on either side) sequence context windows from the African (**Supplementary File 5.1**) and European (**Supplementary File 5.2**) populations from the 1000 Genomes project and found the rates to highly correlated and similar ($R^2 = 0.996$, Pearson correlation = 0.998, **Figure 5.1**).

Next I evaluated the performance of my mutation rate estimates on 4,748 *de novo* events from a high quality pedigree sequencing dataset from 78 Icelandic parent-offspring trios [16]. I compared the observed and the expected *de novo* mutations under different substitution classes, and found that my *de novo* estimates accurately predicted the distribution under each class (**Table 5.1**). In a previous work [29], I discovered several novel mutation promoting motifs at ApT dinucleotides, CAAT and TACG motifs. Here, I also found an enrichment of *de novo* mutations at these motifs and my mutation estimates to accurately predict the distribution under each motif class (**Figure 5.2**, **5.3**). Taken collectively, my analyses demonstrate that a sequence context based approach

using polymorphisms from population level sequencing dataset can be used to accurately estimate the *de novo* mutation probabilities at a fine scale.

My previous work [29], also showed that the larger heptanucleotide ("7-mer", three flanking nucleotides) sequence context around a polymorphic site, explained significantly more variability in genome-wide polymorphism rates as compared to smaller commonly used trinucleotide sequence ("3-mer", one flanking nucleotide) context. Previous estimates of *de novo* mutation rate have been limited to either utilizing the 1mer+CpG [17] (no flanking nucleotide except for CpG) or 3-mer [31, 89] sequence contexts. Therefore, I tested different sequence context models on the *de novo* mutational data from 78 Icelandic family trios. I used a similar likelihood comparison framework from my previous work [29] to evaluate competing local sequence context models. I calculated the likelihood of the observed *de novo* variants from a 1-mer+CpG sequence context model (**Supplementary File 5.3**) and found that the 3-mer model based rates (**Supplementary File 5.4**) significantly improved the fit to the data (log likelihood improvement of 4, **Table 5.2**). I then evaluated if additional local nucleotides improved the fit to the data and found that the 7-mer context based rates further improved the fit to the data (log likelihood improvement of 271, **Table 5.2**) compared to the 3-mer model. Next, I evaluated the performance of 3-mer sequence context based *de novo* mutational rate estimates on mutational data and found that 3-mer model did not accurately or as well as the 7-mer model, predict the observed *de novo* mutations under all the substitutions classes or motif (**Figure 5.2**, **5.3**). Therefore, together these analyses indicate for the first time that 7-mer context based *de novo* mutation rates best explain the observed distribution of germline *de novo* mutations found in humans.

Higher estimate and variability in coding *de novo* mutation rate

Protein coding genes unlike the intergenic noncoding regions, are GC rich [159, 160] and previous studies have reported a higher overall rate of mutation here relative to the rest of the genome [89]. Using my more informative estimates of heptanucleotide sequence context based

de novo mutation rates, I also find an overall higher rate of mutation in the protein coding genes. However, unlike previous estimates [89, 102] of ~$1.5*10^{-8}$ mutations per base pair, per generation in the coding region, I find a slightly higher mutation rate estimate of $1.72*10^{-8}$ ($P << 10^{-6}$). This suggests that the heptanucleotide sequence context based rates capture previously unidentified higher mutation rate causing sequence contexts in the protein coding region. Next, I investigate the reason behind the higher overall mutation rate estimate from a heptanucleotide sequence context model. I find the exonic mutation rate without considering the GC sequence content, because protein coding genes have more GC content which has been shown to result in a higher rate of mutation [159, 160]. Not surprisingly, I report a lower rate of mutation ($1.38*10^{-8}$), more close to the overall genome-wide rate of mutation of $1.2*10^{-8}$ mutations per base pair, per generation. This strongly suggests that the heptanucleotide context based coding mutation rate estimates are higher because they capture more variability in the GC context in the protein coding region. This is also in sync with my previous result [29], where I had demonstrated significant variability within heptanucleotide sequence contexts at CpG sites and reported mutation enriching motif at TACG sequence context.

To further validate my findings of higher rate of mutation identified from heptanucleotide sequence context based estimates, I compared the observed *de novo* mutations in an Icelandic family trio sequencing dataset with expectations from my model. I find that my results match the observed mutations closely, although the expectations from a 1mer+CpG context based coding mutation rate model were also within the 95% confidence interval (Observed mutation 46, Mean of expectation from heptanucleotide context 45.24, 1mer+CpG context 41.38). Since, very few mutations were observed in this trio sequencing dataset to meaningfully compare the heptanucleotide estimates of *de novo* mutation rate with the 1mer+CpG context, I compared the fit of our sequence context models on a larger polymorphism dataset from 6,503 individuals of European and African ancestry [92]. I limited the analysis to recently originated rare variants which have not been subjected to force of selection and mostly shaped due to mutational forces

[94] and found that the 7-mer model explained the patterns of these polymorphisms better (**Table 5.3**) than any other model. Taken together, these results suggest for the first time an even higher rate of mutation in the protein coding region compared to the genome-wide average.

Previous studies have found variation in the *de novo* germline mutation rate between genes owing to different gene specific sequence contexts and length [31]. Here, I find the overall germline mutation rate across all defined transcripts, using the informative heptanucleotide sequence context based rates, and also report significant variation (**Figure 5.4**). Further investigating the role of sequence context in determining this variability, I normalized the gene specific mutation rate with gene length, and also report significant variability (**Figure 5.5**), suggesting that local sequence context differs between genes and results in significant variation in genic mutation rate. Next, I focused on the intra-genic variability in mutation rate. Previous studies have reported the first exon in a eukaryotic gene to gene to have a higher GC content [161]. Since higher GC content can result in higher mutation rate, so I calculated the *de novo* mutation rate within each exon of my defined transcripts. I found the mutation rate to vary significantly across the gene and the normalized rate at first exon to be higher than the remaining exons or even the single exon genes (**Figure 5.6**). This suggests that genes might accrue mutations at different rate (both overall, and at specific sub-genic regions) beyond expectation due to gene length and cautions against standard normalization approaches by only length or simple sequence context models. Moreover, this has direct implication in clinical gene mapping studies, because a gene can have a higher burden of mutation compared to another, owing to different local sequence context.

Issues with other approaches for *de novo* mutation simulation

The community has successfully demonstrated the role of *de novo* mutations in several diseases including Autism [52, 90], Epilepsy [103] and Congenital Heart disease [53], and have found several causal genes with a higher mutation load in affected probands. Virtually all the studies

[90, 156], use a Poisson probability distribution to simulate and compare the observed mutation burden with expected number of *de novo* mutations found from a fixed or a trinucleotide sequence context based rate estimates. While this is informative, this approach makes several assumptions.

First, the Poisson distribution used to find the expected *de novo* mutations underestimates the variance (although the mean of expected mutations is accurate) when predicting the expected mutations after sequencing of haploid chromosome. The correct statistical approach to simulate *de novo* mutations is from a series of independent Bernoulli distribution described in Equation 3

$$Number\ of\ mutations = B(p_1) + B(p_2) + B(p_3) \ldots\ldots + B(p_i) + \ldots B(p_n) \quad (3)$$

Where p_i is the probability of mutation at nucleotide position *i* in the sequence. If we sequence *n* haploid chromosomes, then the actual number of mutations is again described as a sum of independent Bernoulli random variables (Equation 3, repeated *n* time for each haploid chromosome). The commonly used approach to find expected mutations, simplifies Equation 3 to a Poisson distribution in Equation 4.

$$Number\ of\ mutations = Poisson(n \times p)$$

$$p = p_1 + p_2 + p_3 \ldots + p_i \ldots + p_n \quad (4)$$

where *p* is the sum of individual mutation probabilities at each nucleotide position in the sequence. The mean of Equation 5 and 4 is the same as *np*. However, the variance in Equation 3, under the assumption of independence $np - np^2$, while that of a Poisson approximation in Equation 4 is still *np*. This difference of variance is small because individual probability of mutation is very small. However, it must be acknowledged when using Poisson approximation to simulate and predict mutations at hyper-mutable sites or in species with higher background mutation rate.

Second, the Poisson distribution approach to find the expected mutations in different substitution classes (missense, nonsense, deleterious, synonymous etc.), incorrectly estimates the variance in each class (although the mean of expected mutations is accurate). Each nucleotide can mutate to another resulting in a different type of change based on the position in the codon frame. For example, the nucleotide C in codon CGA can mutate to A, resulting in a synonymous change, mutate to G, resulting in a missense change and mutate to T, resulting in a nonsense change. However, these mutation probabilities are not independent of each other and we need to model the covariance. In Equation 5, I describe a scenario where one is interested in estimating the total deleterious mutations over a particular coding region after sequencing *n* haploid chromosomes. The correct statistical approach includes modeling the probability of missense and nonsense mutation at each nucleotide position using a Bernoulli distribution.

$$Number\ of\ deleterious\ mutations = B(p_{missense\ 1}) + B(p_{nonsense\ 1}) + B(p_{missense\ 2}) + B(p_{nonsense\ 2}) \ldots . + B(p_{missense\ n}) + B(p_{nonsense\ n}) \ldots repeated\ for\ each\ chromosome \quad (5)$$

However, the Poisson approach models the number of deleterious mutation as

$$Number\ of\ deleterious\ mutations = Poisson(n \times p_{missense}) + Poisson(n \times p_{nonsense})$$

$$p_{missense} = p_{missense\ 1} + p_{missense\ 2} + \cdots p_{missense\ n}$$

$$p_{nonsense} = p_{nonsense\ 1} + p_{nonsense\ 2} + \cdots p_{nonsense\ n} \quad (6)$$

Where $p_{missense}$ or $p_{nonsense}$ is the sum of missense or nonsense mutation probability at each nucleotide in the sequence. The mean of Equation 5 and 6 is the same as $n*(p_{missense} + p_{nonsense})$. However, the variance in Equation 6 under the Poisson model is $n*(p_{missense} + p_{nonsense})$, while in Equation 5 we have covariance terms because having a missense or a nonsense mutational event at a nucleotide position are not independent of each other. Moreover, the Poisson model further does not include the square terms of missense and nonsense probabilities. While the

covariance and square terms are small (because of small missense and nonsense probabilities), they must be acknowledged when using the Posisson approach.

Together, these results highlight several approximations in existing approaches for simulation and then testing for *de novo* mutation burden over genomic regions. While the Poisson approach, results in accurate estimate of the mean number of expected mutations, the variance can be off when the background rate of mutation is high. This can have implications in mutation burden tests because simulations from Poisson distribution are used to compare the observed burden with expected, and infer significance. While existing studies have mostly focused on a few genes after sequencing of some individuals, studies of future with complex analysis on a large dataset, in hyper-mutable regions or non-human species with higher mutation rate, can be compromised with the Poisson distribution approach.

Toolkit for simulating, distributing and interpreting *de novo* mutations

My central objective here is to develop a framework to find the expected number of mutations over any region of the genome based on a background model of mutation rate, which can then be used for understanding the process of mutagenesis under different conditions [162–164] or finding the causal gene or locus from disease sequencing studies [6, 90, 97]. To achieve this objective, I need (a) an accurate model of variation in background mutation rate and (b) an algorithm to simulate and distribute mutations over any region of the genome. I have previously demonstrated that a heptanucleotide sequence context model, accurately captures the variability in the genome-wide rate of mutation. Therefore, here I use the informative *de novo* mutation rate estimates from the 7-mer sequence context model in my framework. I perform my analysis over any genomic region, which can be specified via genomic coordinates or using ENSEMBL transcript ids. The final output, includes the expected count of mutations over the genomic region, further sub-classified into substitution classes (transition or transversion, if genomic region includes coding territory then synonymous, missense, nonsense, splice, individual amino acid

annotations). I also output the genomic regions with simulated mutations at different positions, which the user can then analyze based on their specifications.

I first develop an algorithm to simulate *de novo* mutations of any genomic region of interest. This is geared towards a user interested in finding the occurrence of *de novo* mutations over a region, after sequencing *n* individuals (or *2n* haploid chromosomes). A common question asked in most family trio disease sequencing studies, is if the observed count of mutations in a gene or a pathway is unusually higher or lower than expected under a background model of mutation rate. An unusual observation will suggest a protective or causal role of the gene or pathway in disease pathogenesis. My algorithm extends the multinomial probability distribution framework, where I simulate mutations at all occurrences of heptanucleotide sequence contexts in a region using the heptanucleotide *de novo* mutation rate estimates. Over multiple simulations and total haploid chromosomes (both given as input by the user) sequenced, I use my algorithm to decide where mutations occur in the genomic region, in each simulation. Finally, I output the 95% confidence interval of the total and substitution class specific mutation count from the simulated data.

Next, I developed an algorithm to distribute *de novo* mutations across any genomic region, in order to generate an expected count of mutations (with variance) at all positions in the region. This is geared towards a user interested in comparing the mutations observed in a disease ascertained sample, with an expected distribution from a background model of mutation rate. A common question asked after causal gene discovery is, whether mutations in a disease ascertained sample are enriched at certain sub-genic sequences. An enrichment of mutations at a particular genic locus in a disease ascertained sample can suggest higher pathogenicity, and vice versa for mutations observed in a sample ascertained against disease status. Here too, I extend my multinomial probability distribution framework to distribute mutations over multiple simulations (both given as input by the user) on a particular genomic region. I find the probability of mutation at each position in the genomic region, and then use a multinomial distribution with the calculated probabilities to distribute the mutations. Finally, I output the 95% confidence

interval of substitution class specific mutation count from the multiple simulations used to distribute mutations over the genomic region.

Together, these functionalities of simulation and distribution provide a systematic framework to find the expected spectrum of mutations over a genomic region, based on the background mutation rate. While previous approaches mostly focused on the simulation aspect [156], using the less informative trinucleotide sequence context based mutation estimates, my approach including the novel distribution algorithm provides a generalized framework for interpretation of *de novo* mutations in clinical studies. Moreover, my approach also improves the resolution by finding expected mutations over any genomic region (gene set or sub-genic loci) and for any substitution class. I implement these functionalities as a part of a toolkit called SimDenovo, which can serve as a useful resource for the community interested in analyzing and interpreting disease sequencing mutational data.

CHAPTER 6: Discussion and Future work

Mutation is the most important force that has shaped our genomes since we evolved from single cell species [1]. It generates genetic variation on which evolution acts [2], causes variation between individuals of same [3] and different species [4], results in cell to cell heterogeneity [5] and is responsible for genetic disorders; inherited [6] or somatic [7] like cancer. Hence, understanding how mutations originate across the genome is fundamental to our understanding of life. Previous work over the last century, has successfully discovered and described the process of mutation [3, 8] in single celled organisms to complex eukaryotic species like *Homo sapiens*. An important finding across the entire evolutionary tree has been that the rate of mutation varies significantly across genomes [28, 30], and it has critical implications in understanding evolution, disease and mutagenesis biology.

Previous studies have shown that local sequence context correlates with, and in specific cases, directly modifies the rate of mutation [30]. An exquisite example is the CpG sequence context, where spontaneous deamination of methylated cytosine elevates the cytosine to thymine mutation rate by ~15-fold higher relative to the genome-wide background [28, 104]. The community has incorporated local sequence context around the polymorphic site, *i.e.* 1 base pair on either 5' or 3' end for a total of 3 nucleotides or trinucleotide sequence context, to model and understand the variability in mutation rate [30–32, 89]. However, the role of larger sequence context in explaining the variation in mutation rate has not been systematically evaluated. Therefore, in this thesis, (a) I proposed a statistical model for the probability of nucleotide substitution in the genome based on windows of nucleotide sequence context, (b) using this model, proposed a framework to statistically test competing windows of sequence context, and concluded that a broad (heptanucleotide) context best explained population level and *de novo* mutational data, (c) utilized inferences about the rates generated from my predicted sequence context model to ask several questions, ranging from identification of causal genes, testing

specific disease hypotheses, or sub-gene burden of de novo mutations, as applied to multiple disease sequencing datasets.

In **chapter 2**, I evaluated different sequence context models found that a larger heptanucleotide sequence context model best explained patterns of nucleotide substitution observed in the human genome. I demonstrate that the commonly used context that includes one nucleotide flanking a polymorphic site does not fully capture the complete spectrum of where, what type, and how frequently nucleotides are expected to change. I also find novel variability in substitution probabilities at CpG sites, and found certain heptanucleotide CpG contexts to have a much higher or lower rate of substitution. Moreover, I also show that this previously unidentified variability at CpG sites cannot be fully explained by differential methylation intensity patterns. Furthermore, I identify novel mutation promoting motifs at TACG and ApT and CAAT sites. Finally, I demonstrate that nucleotide substitution probabilities capture features of germline mutation rate as they are consistent and correlated (a) across the frequency spectrum of variants, (b) between high and low recombination regions, (c) with human primate divergence and (d) with real *de novo* mutational counts.

One question in the field has been how much sequence context can explain patterns of nucleotide substitution in genomes [165]. My results suggest that a substantial fraction can be robustly predicted by sequence context alone, although specific substitution classes may require more features than just sequence context. I also acknowledge that context models beyond three flanking nucleotides were not considered. The regression approach that I presented does suggest that the 7-mer models could be refined, perhaps allowing broader context to be considered.

Furthermore, evolutionary genetics studies require an estimate of mutation rate for accurately dating divergence events [166]. Since, most of these studies incorporate a simplistic model of mutation rate variation at CpG and non CpG sites, they can vastly benefit from a more informative model of mutation rate variation at the level of sequence context.

While I did not apply my model to other species, the strong correlation with divergence suggests that the features of mutation are potentially conserved across primates. The same framework can be applied to population level sequencing datasets in other species [167, 168] to discover features of mutation (if any) specific to them. Moreover, comparative genomics applications to identify non-neutrally evolving regions, genome alignments, or tree reconstruction [157], would benefit from my accurate model of nucleotide substitution.

Here, by using population level data with millions of SNPs rather than smaller *de novo* mutations dataset, I was able to improve the resolution of substitution models and model variability in substitution/mutation rates. Because SNPs in intergenic, noncoding regions are mostly shaped due to forces of mutation (and population demographic forces), this strategy can be applied on larger datasets (*e.g.,* dataset from sequencing of 100K individuals from UK) to further improve the human substitution probability estimates and learn about features of germline mutation. I also acknowledge that a number of features remain to be formally evaluated in the genome [8], for example, recombination in the coding genome [169] or replication timing [170]. My framework has the flexibility to model the complexity found in any sequences that contain features hypothesized to be important.

With an appropriate background model for nucleotide substitution, novel statistics for clinical re-sequencing studies can also be envisioned, based on the occurrence of discovered variation. Such approaches may complement statistics that assay allele frequency differences between cases and controls at one or more polymorphic sites. While the underlying mechanisms that determine how nucleotide sequences change over time remain to be addressed, I posit that features identified from my model provide important clues in elucidating these fundamental principles.

In **chapter 3**, I extended my substitution probability framework to the coding region and characterized average selective pressures operating in the coding genome at a finer level of

detail. I first demonstrate that a heptanucleotide sequence context model that incorporates both (a) the mutational forces using substitution rates in the noncoding region, and (b) selective forces through the codon position effects in the coding region, best explain patterns of coding substitutions compared to a commonly used trinucleotide sequence context model. My model also indicates substantial variability across all amino acid replacement classes, and, in some cases, synonymous substitutions that were less prone to change than missense or even nonsense substitutions. Furthermore, I modeled the average selective force acting on the coding genome at each sequence context stratified by codon position, and developed several clinical utilities of interest consistent with purifying selection and disease hypothesis [47]. First, I show that nonsense and missense variants in 1KG dataset of population controls have a higher selective constraint on them compared to synonymous polymorphisms, but less than on the putative disease causing variants in the HGMD dataset. Second, I develop a statistic to quantify the functional intolerance or selective constraint at the level of a gene. Referred to in the community as a “gene score” [31, 171], I demonstrate that such measures have the best performance to grade intolerance when incorporating heptanucleotide sequence context. I show that genes associated with essentiality in humans or are ubiquitously expressed in different tissues, and associated with several neuropsychiatric diseases have more intolerant gene scores, while those associated with immune function are more tolerant. Third, I further improve the resolution and find amino acid scores that measure the intolerance of an amino acid to functional changes and find several examples of localized intolerance at the level of amino acid but tolerance at overall genic level. Finally, I apply these scores to a *de novo* disease sequencing dataset of Autism probands, and show that previously discovered genes and variants in this set of genes have more intolerance and further suggest the role of some novel genes in Autism.

Here, I modeled the selective constraint acting in the coding genome at a fine scale, by comparing the polymorphisms in the coding region with that in the intergenic noncoding region. Assuming that noncoding regions I curated are shaped mostly due to mutational forces

(attempting to minimize selective pressures), I utilized the ratio of the two substitution probabilities to find the selective force at each sequence context in the coding region to quantify the extent of purifying selection of different types of mutation. In future, with large exome sequencing datasets in the coding region, the coding substitution probability estimates can be further improved to obtain more accurate estimates of selection. Moreover, amino acid and gene scores can be further improved as in the current dataset of 1KG controls, some genes or amino acids had very few substitutions to train our model on. Furthermore, with larger datasets the resolution of intolerance scores can be further improved and can also be developed for other annotations like exon level or different sub-genic locus.

I also find significant variation in substitution patterns in the coding region, and found some amino acid substitutions (even nonsense changes like from Arginine to Stop) to occur more than others (synonymous or missense changes). Therefore, more powerful rare variant burden tests can also be envisioned which incorporate the background probability of occurrence of a rare variant in the sequencing dataset.

While the community has been extensively using intolerance scores for prioritizing genes discovered from clinical sequencing datasets [31, 125], my intolerance scores can further benefit such studies. Moreover, my unique amino acid scores, which measure intolerance at the level of amino acid can be further used to prioritize both genes and variants discovered from disease sequencing datasets for follow up. Finally, the studies that infer the presence and strength of selection on genes might further benefit by incorporating my intolerance scores.

In **chapter 4**, I developed a generalized approach, which models the variability in mutational probabilities at sequence context level and finds enrichment of *de novo* mutations at sub-genic loci from sequencing datasets. Enrichment of mutations at a locus in disease ascertained samples can suggests localized pathogenicity within a gene, and can be used to further elucidate disease mechanisms. I applied the generalized approach to a large dataset of *de novo* germline

mutations in bilateral RB patients without a previous family history of disease. First, I show that the frequency of nonsense mutations at CpG sites is compatible with the background model for the known, elevated rate of mutation at these sites. A parsimonious interpretation of this result is simply that nonsense mutations at CpG sites in RB1 are, in fact, not preferentially RB pathogenic. Instead, the abundance of Arginine to Stop mutations can simply be explained by (i) ascertainment of RB affected probands, (ii) that LoF at RB1 causes RB, and (iii) the mutability of this sequence context [14, 33]. Second, I identified heterogeneity in the frequency of essential donor splice-site mutations across RB1. In particular, I found a depletion of essential donor splice site mutations in intron 5, explainable by the fact that exon 5 skipping retains the coding frame (at the loss of a 13 amino acid deletion) and thus may only be weakly penetrant. I also found more mutations at essential donor splice-sites of introns 6 and 12 than predicted by my model, which result in frame-shift and putative loss of function (LoF) mutations. Here, I note that essential donor splice-sites in other introns also result in frame-shift and putative LoF. Thus, a mechanistic explanation as to why exon 6 and 12 skipping and consequent frame-shift LoF would be specifically ascertained in our probands remains elusive. Nonetheless, statistical quantification of this specific enrichment, to my knowledge, has not been previously reported. Finally, I quantified the excess of missense mutations in Exon 20, localized specifically to Arg661Trp. While I noted the recurrence of five mutations to this specific codon, as well as and enrichment in another LOVD dataset, I was not able to distinguish the relative frequency of this mutation from the rate of nonsense owing to the small number of events in the dataset. Previous reports in the literature gives some indication that this mutation is indeed low penetrance [27–29], and my results are consistent with these reports.

A major challenge in de novo mutational studies of rare and complex disease is to not only identify new pathogenic mutations, but also to statistically quantitate the enrichment of specific types of pathogenic mutations within a gene, in order to improve the understanding of gene-specific disease etiology. Here, my motivation was based on the need to statistically evaluate

specific hypothesis about the relative abundance – and inference about pathogenicity – of *de novo* mutations identified in probands selected for bilateral RB. This study of sporadic RB cases identified under a research protocol represents the single largest dataset of de novo mutations in the RB1 gene reported to date. Thus, it removes many uncertainties associated with other data sets where there are many sources of non-homogeneity including sample ascertainment and methods used for mutation detection. Moreover, the significance of identifying *de novo* mutations for affected probands includes not only clinical management decisions, but also risk of a second cancer in the future as well as having additional, affected offspring. Thus, investigating the pathogenicity of *de novo* mutations in this study is both mechanistically and clinically relevant. However, this analysis using my generalized framework would vastly benefit from a larger dataset of mutations and can allow for testing and discovery of several novel disease hypotheses.

Moreover, with sufficient data and a specific probabilistic model, it is conceivable to utilize my approach to derive posterior distributions for penetrance for different classes of mutations. Such may be the focus of future work. I focused here exclusively on the analysis of RB, owing to the systematic extent that this disease has been previously studied, the preponderance of existing data sets, and minimal genetic heterogeneity for the condition. Despite this, my efforts helped to clarify existing hypotheses in the field around mutational mechanisms for the gene and point to new areas to study for this already well-studied disease. That said, my framework could be readily applied to other Mendelian diseases or complex disorders. While each disease endpoint will have particular biological mechanisms to elucidate, the model and approach I present should provide a statistical framework to identify sequence-based features that point to unknown mechanisms underlying human disease.

In **chapter 5**, I generated accurate and informative estimates of *de novo* mutation rate for different sequence context models. I showed that estimates of *de novo* mutation rate at heptanucleotide sequence contexts are more informative than the commonly used trinucleotide estimates for contexts on an external dataset of *de novo* mutations. Moreover, the

heptanucleotide estimates accurately predict the observed pattern of mutations. I also find a higher yet accurate estimate of the mutation rate in the coding region suggesting the presence of more mutable GC rich contexts, which is captured by the heptanucleotide sequence context. I also demonstrate significant inter and intra-genic variability in mutation rate across the genes, suggesting the need and importance of a mutation rate framework which can incorporate this variability and interpret mutations in a disease sequencing study. I show that the commonly used approach for simulating mutations is informative, but uses many assumptions which must be acknowledged in more complex studies. Finally, I develop a generalized framework which can (a) find expected mutations over any genomic region after sequencing *n* individuals and (b) distribute *N* mutations and find expected at any context under a background model for mutation rate. Together these functionalities can be used to simulate, distribute and interpret *de novo* mutations in different diseases.

Here, I was motivated by the need of the community to have an accurate estimate of *de novo* germline mutation rate which can then be used in several applications from modeling of evolutionary processes [15] to finding the causal gene from de novo sequencing studies for different diseases [52, 90]. Since, I had earlier shown that the heptanucleotide sequence context captures feature of germline mutation rate, so now I find actual estimates at each context, such that the overall *de novo* mutation rate is fixed at $1.2*10^{-8}$ mutations per nucleotide, per generation. In future, if the overall mutation rate gets revised to a higher or lower number, or if we want to find it in different species, we can correspondingly scale the mutation rate estimates at each sequence context. I also validate my mutation rate estimates on an external dataset of ~4748 mutations, but it would be interesting to exhaustively compare the accuracy on larger datasets.

I also find a higher mutation rate estimate in the protein coding region. Protein coding genes have a higher GC content, which are more mutable sequences and hence previous studies had also reported a higher rate of mutation in the coding region [89, 90]. However, I find an even higher

mutation rate, suggesting the presence of some more mutable contexts. Further improving the resolution of these mutable contexts can help clarify evolutionary processes and mechanism that resulted in more mutable coding regions. Moreover, it would be interesting to characterize the variability in gene specific mutation rate (after normalizing by length), and to understand the complex relationship between higher mutation rate at a gene and overall fitness [172]. I also demonstrate that the first exon has a higher mutation rate than the rest of the gene body, but the evolutionary processes shaping this phenomenon are not clear. Since higher mutation rate can result in more deleterious mutations, which have a detrimental effect on fitness [173], future work could focus on unraveling the beneficial effects to organismal fitness from having a higher localized mutation rate within a gene.

Finally, using my mutation rate framework for simulation and distribution of mutations, powerful tests can be envisioned that measure the burden of mutation rate in several clinical studies. Current approaches [90, 156, 174] only compare the burden of mutations in cases and controls. While this is informative, it does not capture other features which can add more power to the testing strategy. For example, using a fine scale model of mutation rate variability and my unique distribution function, one can develop new test that also compares the deleterious to non-deleterious mutation rate burden in cases and controls. This additional feature, combined with other features of localized or specific higher burden of mutation rate can be used to compare the mutational burden and test for specific hypotheses.

In conclusion, the general theme of my thesis research has been to characterize and model the variability in human mutation rate. Mutation rare is a fundamental quotient in human genetics, and its accurately modeling can be used to answer innumerable questions ranging from basic biology to translational therapeutic applications. The work presented in my dissertation has laid a solid foundation for study of variation in mutation rate, and provides a systematic framework for interpretation of clinical sequencing data to find causal genes or test for specific disease mechanisms. Studies of future can build on my mutation rate framework to advance our

knowledge of human origins and evolutionary processes, and also to summarize and interpret the vast disease sequencing datasets for all genetic diseases.

TABLES

Table 2.1 Comparison between substitution probabilities from HapMap and1KG data. R^2 and correlation between the substitution probabilities estimated using HapMap and 1000 Genomes variant data from the intergenic non-coding genome, for different sequence context models (3-mer model with randomized sequence context beyond adjacent nucleotides, 7-mer model). Also shown is the comparison specific for CpG and nonCpG sequence contexts.

Sequence Context type	R^2 of 7-mer model	Correlation with 7-mer model	R^2 of 3-mer model with randomized sequence context beyond adjacent nucleotides	Correlation of 3-mer model with randomized sequence context beyond adjacent nucleotides
All	0.91	0.95	0.84	0.91
CpG only contexts	0.88	0.94	0.78	0.88
non CpG contexts	0.75	0.87	0.6	0.77

Table 2.2 Substitution probabilities for different populations and chromosomes. Average nucleotide substitution probabilities for different population groups (African, European, and Asian) on different types of regions (coding versus intergenic non-coding) and on different chromosomes (All autosomal versus X chromosome).

	African Substitution Probability	Asian Substitution Probability	European Substitution Probability
All IGR	0.009454398	0.005240961	0.006142976
Autosomal IGR	0.00971703	0.00541551	0.006339118
X chromosome IGR	0.006122926	0.00302682	0.003654918

Autosomal Coding	0.007195663	0.004923134	0.005478647

Table 2.3 Variance in a class explained by different models. Summary of Summary and performance of forward regression model for feature selection using the 7-mer context in the intergenic non-coding genome. % Substitutions represents the percentage of substitutions for that class observed in the genome. # Parameters represents the number of features selected in the best 7-mer model. Model R^2 (7-mer) reflects prediction accuracy in the test dataset alone (not used for model training) with the best model using heptanucleotide sequence context features. Model R^2 (3-mer) denotes the prediction accuracy with only trinucleotide sequence context features.

Substitution Class	# Contexts	% Substitutions	# Parameters	Model R^2 (7-mer)	Model R^2 (3-mer)
Outside CpG Dinucleotide Context					
A-to-C	4096	7.3	266	56.5	11.2
A-to-G	4096	28.2	366	91.5	40.9
A-to-T	4096	7.1	197	58.7	37.4
C-to-A	3072	8.5	282	83.5	30.0
C-to-G	3072	7.5	268	81.0	17.1
C-to-T	3072	24.4	254	86.8	37.6
Within CpG Dinucleotide Context					
C to A	1024	1.0	26	58.3	19.0

C to G	1024	0.8	95	48.7	9.5
C to T	1024	15.2	96	93.1	44.4

Table 2.4 Sequence motifs identified from substitution probabilities from 1KG data. Enrichment of motifs identified in posterior nucleotide substitution probabilities for the 7-mer sequence context models inferred from intergenic non-coding genome. CpG+ indicates the distribution of sequence contexts which include a CpG site (4th position polymorphic site is C, 5th position fixed as G). Enrichment P-value is based on the enrichment of the motif in the 1% tail of the given substitution class: “Higher” implies enrichment in the upper 1% tail of the sequence context probability distribution, “Lower” implies enrichment in the lower 1% tail. Odds ratio and [95% CI] denotes the odds ratio (and 95% confidence interval) of enrichment of motif in the upper or lower 1% tail of the sequence context probability distribution. Fold change in substitution rate denotes the fold increase or decrease in substitution rates for the motif relative to its substitution class.

Motif	Substitution Class	Effect on Substitution Probability	Enrichment P-value	Odds ratio and [95% CI]	Fold change in substitution rate
NNN**C**GNN	C-to-T	Higher	2×10^{-26}	134.4 [18.4 - 977.4]	13.9
	C-to-G	Higher	1×10^{-13}	12.8 [5.9 - 27.7]	2.4
	C-to-A	Higher	9×10^{-22}	60.8 [14.6 - 252.1]	2.7

N[A/C/G][C/G/T]**C**G CG	C-to-T (CpG+)	Lower	7×10^{-16}	366.3 [45.6 - 2939.5]	1.5
Poly T and Poly A combination (AAA**A**TTT, TTT**A**AAA)	A-to-T	Higher	9×10^{-5}	304.2 [31.0 - 2987.6]	12.7
Quad A (AAA**A**NNN, NAA**A**ANN, NNA**A**AAN,NNN**A**A AA)	A-to-G	Lower	5×10^{-10}	10.2 [7.3 - 14.1]	1.9
NTA**C**G[C/G][A/C/G]	C-to-T (CpG+)	Higher	1×10^{-10}	102.5 [27.4 - 383.2]	1.7
NNT**A**CGN	A-to-C	Lower	3×10^{-4}	9.4 [3.6 - 24.8]	1.5
NNN**A**TNN	A-to-T	Higher	2×10^{-17}	22.3 [8.7 - 57.1]	1.6
	A-to-G	Higher	1×10^{-25}	131.2 [18.0 - 954.2]	2.0
[C/T]CA**A**T[C/G/T]N	A-to-G	Higher	8×10^{-53}	5966 [2091 - 17021]	5.1

Table 2.5 Sequence motifs identified from substitution probabilities from HapMap data. Enrichment of motifs identified in nucleotide substitution probabilities inferred from HapMap variant data in the intergenic non-coding genome. CpG+ indicates the distribution of sequence contexts which include a CpG site (4th position polymorphic site is C, 5th position fixed as G). Enrichment P-value is based on the enrichment of the motif in the 1% tail of the given substitution

class: “Higher” implies enrichment in the upper 1% tail of the sequence context probability distribution, “Lower” implies enrichment in the lower 1% tail. Odds ratio and [95% CI] denotes the odds ratio (and 95% confidence interval) of enrichment of motif in the upper or lower 1% tail of the sequence context probability distribution. Fold change in substitution rate denotes the fold increase or decrease in substitution rates for the motif relative to its substitution class.

Motif	Substitution Class	Effect on Substitution Probability	Enrichment P-value	Odd’s ratio and [95% CI]	Fold difference in substitution rate
NNN**C**GNN	C-to-T	Higher	2×10^{-25}	134.4 [18.5 - 977.4]	10.1
	C-to-G	Higher	2×10^{-26}	131.4 [18.1 - 956.2]	2.8
	C-to-A	Higher	1×10^{-25}	128.1 [17.6 - 933.0]	3
N[A/C/G][C/G/T] **C**GCG	C-to-T (CpG+)	Lower	2×10^{-16}	43.6 [12.2 - 155.3]	4.4
Poly T and Poly A combination (AAA**A**TTT, TTT**A**AAA)	A-to-T	Higher	9×10^{-5}	304.2 [31.0 - 2987.6]	7.7
Quad A (AAA**A**NNN, NAA**A**ANN, NNA**A**AAN,NNN **A**AAA)	A-to-G	Lower	1×10^{-2}	3.1 [2.1 - 4.8]	2.1
NTA**C**G[C/G][A/C /G]	C-to-T (CpG+)	Higher	2×10^{-6}	43.6 [12.2 - 155.3]	2.3
NNT**A**CGN	A-to-C	NA	Not Significant	NA	1.1
NNN**A**TNN	A-to-T	NA	Not Significant	NA	1.5
	A-to-G	Higher	2×10^{-21}	41.5 [12.8 - 134.5]	1.8

[C/T]CA**A**T[C/G/T]N	A-to-G	Higher	8×10^{-53}	5966 [2091 - 17021]	4.4

Table 2.6 Substitution probabilities and human primate divergence. R^2 and correlation between the substitution probabilities in intergenic non-coding genome and human primate divergence at a context, for different sequence models (3-mer model with randomized sequence context beyond adjacent nucleotides, 7-mer model). Also shown is the comparison specific for CpG and nonCpG sequence contexts.

CpG Contexts		
Correlation with 7-mer model	R^2 of 3-mer model with randomized sequence context beyond adjacent nucleotides	Correlation of 3-mer model with randomized sequence context beyond adjacent nucleotides
0.99	0.92	0.96
0.92	0.44	0.67
0.69	0.16	0.41
Non CpG Contexts		
Correlation with 7-mer model	R^2 of 3-mer model with randomized sequence context beyond adjacent nucleotides	Correlation of 3-mer model with randomized sequence context beyond adjacent nucleotides
0.98	0.93	0.96
0.91	0.41	0.64
0.72	0.22	0.47

Table 2.7 De novo mutations at identified motifs. Expected (95% CI) and observed *de novo* mutations for each class of change calculated on high quality pedigree sequencing data from 78 trios[16]. If number of observed mutations fall in the expected confidence interval, then I denote it “As expected” otherwise as “Higher than expected”.

Mutation Class	95% CI of expected *de novo* mutations	Observed *de novo* mutations	Comparison of observed mutations with expected mutations
C-to-T	[1804, 1974]	1,952	As expected
C-to-T at **C**G sites	[67, 104]	816	Higher than expected
C-to-T at TA**C**G sites	[1, 8]	54	Higher than expected
A-to-G	[1245, 1388]	1,279	As expected
A-to-G at **A**T sites	[307, 380]	588	Higher than expected
A-to-G at CA**A**T sites	[11, 28]	72	Higher than expected
A-to-G at CT**A**T sites	[9, 24]	67	Higher than expected
A-to-T	[303, 375]	316	As expected
A-to-T at **A**T sites	[70, 107]	120	Higher than expected

Table 3.1 Comparison of different substitution probability models in the coding region. Natural logarithm of the approximate Bayes Factor comparing the posterior likelihoods of the 3-mer context model with and without accounting for codon context, and my proposed 7-mer context model which does include codon context on multiple data sets. I present data from the African and European groups (1KG) and an analysis of the EVS dataset (individuals of European ancestry) after filtering out variants with minor allele frequencies less than 0.03%.

Dataset	Population	Approximate ln(Bayes Factor) comparing the basic 3-mer model with the coding 7-mer model	Approximate ln(Bayes Factor) comparing the basic 3-mer model with codon offset with the coding 7-mer model
1000 Genomes	African	599417	529074
1000 Genomes	European	611063	546810

EVS	European	568402	504888

Model	Number of Parameters
Basic 3-mer model	192
Basic 3-mer model with codon offset	576
7-mer coding model	73728

Table 3.2 Comparison of gene tolerance scores from different approaches. Prediction accuracy of gene tolerance scores to classify membership in various gene sets analyzed in this study. Area under the curve (AUC) calculations for gene scores of and my 7-mer codon context gene scores.

Gene Classes	AUC of Petrovaski	AUC of Samocha	AUC of Aggarwala
Essential	0.66	0.55	0.68
Ubiquitous	0.59	0.53	0.7
Immune	0.53	0.51	0.54
Omim Denovo	0.66	0.58	0.67
Omim Dominant	0.65	0.54	0.65
Omim Haploinsufficient	0.73	0.6	0.73
Olfactory	0.81	0.52	0.83
Keratin	0.69	0.51	0.7
Autism	0.83	0.67	0.82
Intellectual Disability	0.82	0.73	0.89
Developmental Disorder	0.81	0.7	0.8
Epilepsy	0.92	0.81	0.91
All Disease (Autism, Epilepsy, Intellectual Disability and Developmental Disorder)	0.82	0.7	0.86

Table 4.1 Variant counts in different categories from RB and ExAC datasets. Counts of *de novo* mutations in *RB1* ascertained from RB patients, and singleton variants identified in ExAC from (non-Finnish) Europeans for various subtypes.

Variant Type	RB *de novo* mutations	ExAC singletons
Overall	268	149
Nonsense	150	1
Missense	27	56
Essential Splice	86	1
Intronic	5	91

Table 4.2 Enrichment of nonsense mutations at different amino acids. Comparison of the observed number of nonsense *de novo* mutations to the simulated frequency predicted by my sequence context model. Data shown for all amino acids which can change to a stop codon as well as Arginine codon partitioned by CpG context. CI: Confidence Interval.

Amino Acid	99% CI of simulation	Observed variants	Empirical P
Lysine	[0, 11]	3	0.336
Serine	[2, 15]	6	0.404
Leucine	[1, 13]	5	0.454
Glutamine	[5, 23]	15	0.385
Tryptophan	[1, 13]	3	0.126
Arginine	[73, 104]	95	0.188
Glutamic	[4, 20]	14	0.243

Glycine	[0, 6]	3	0.211
Cysteine	[0, 7]	1	0.399
Tyrosine	[2, 16]	5	0.143

Arginine Codon	99% CI of simulation	Observed variants	Empirical P
CGA	[73, 104]	93	0.237
AGA	[0, 4]	2	0.209

Table 4.3 Enrichment of nonsense mutations at different exons in an external dataset. Comparison of observed mutations and the simulated frequency of nonsense changes per exon, to find differential pathogenicity within nonsense mutations. Analysis was performed on data from Onadim and Houdayer groups to further validate my results on an external dataset.

Exon	99%CI of Simulation	Observed variants	P value
2	[0, 8]	3	0.536
3	[0, 7]	1	0.328
4	[0, 5]	2	0.392
5	[0, 3]	2	0.187
6	[0, 4]	0	0.365
7	[0, 6]	1	0.379
8	[3, 17]	8	0.374
9	[0, 3]	0	0.694
10	[2, 15]	7	0.556
11	[1, 13]	12	0.015
12	[0, 7]	1	0.264
13	[0, 7]	1	0.319

14	[4, 18]	13	0.233
15	[1, 13]	9	0.17
16	[0, 4]	1	0.572
17	[5, 21]	16	0.149
18	[1, 13]	4	0.264
19	[0, 6]	3	0.365
20	[0, 8]	3	0.508
21	[0, 5]	0	0.275
22	[0, 6]	5	0.054
23	[1, 12]	8	0.21
24	[0, 4]	0	0.474
25	[0, 6]	0	0.241
26	[0, 3]	0	0.452
27	**[3, 18]**	**0**	**0**

Table 4.4 Enrichment of nonsense mutations at different AA's in an external dataset. Comparison of observed mutations and the simulated frequency of nonsense changes to find differential pathogenicity within nonsense mutations. Data shown for all amino acids and two arginine codons (99% CI) which can change to a stop codon. Analysis was performed on data from Onadim and Houdayer groups to further validate my results on an external dataset.

Amino Acid	99% CI of simulation	Observed variants	P value
Lysine	[0, 8]	1	0.179
Serine	[1, 12]	4	0.491
Leucine	[0, 10]	2	0.246
Glutamine	[3, 17]	9	0.554

Tryptophan	[0, 10]	3	0.424
Arginine	[48, 72]	66	0.107
Glutamic	[2, 15]	9	0.318
Glycine	[0, 4]	0	0.343
Cysteine	[0, 5]	1	0.6
Tyrosine	[1, 12]	5	0.509
Arginine Codon	99% CI of simulation	Observed variants	P value
CGA	[47, 71]	63	0.241
AGA	[0, 3]	3	0.016

Table 4.5 Unusual distribution of essential donor splice mutations at different exons**.** Comparison of the observed number of essential donor splice-site *de novo* mutations at exons 6, 12, and 5 to the simulated frequency predicted by my sequence context model. “LOVD count” denotes the point variants observed at this site in the LOVD dataset. CI: Confidence Interval.

Location	99% CI of simulation	Observed variants	Empirical P	LOVD count
Exon 6	[0, 2]	3	3×10^{-4}	40
Exon 6	[0, 4]	9	$< 10^{-6}$	40
Exon 12	[0, 10]	13	4×10^{-4}	67
Exon 5	[1, 12]	0	3×10^{-3}	2

Table 4.6 Polyphen prediction on different missense variants. Polyphen predictions on the *de novo* germline missense mutations or some potential variants near codon 661 in *RB1* gene. “Polyphen2_format” is the variant format accepted by the Polyphen2 tool. “Polyphen_prediction” is the result of Polyphen2 on the missense variant.

All *de novo* germline missense mutations				
gDNA position	Reference allele	Alternate allele	Polyphen2 format	Polyphen2 prediction
39554	G	A	chr13:48916843 G/A	benign
77027	A	G	chr13:48954327 A/G	probably damaging
156836	C	A	chr13:49033967 C/A	probably damaging
76454	T	G	chr13:48953754 T/G	probably damaging
78278	C	T	chr13:48955578 C/T	probably damaging
156717	T	C	chr13:49033848 T/C	probably damaging
156713	C	T	chr13:49033844 C/T	probably damaging
156713	C	T	chr13:49033844 C/T	probably damaging
156713	C	T	chr13:49033844 C/T	probably damaging
156713	C	T	chr13:49033844 C/T	probably damaging
156713	C	T	chr13:49033844 C/T	probably damaging
73868	A	T	chr13:48951169 A/T	probably damaging
59789	A	G	chr13:48937089 A/G	probably damaging
160740	G	A	chr13:49037877 G/A	probably damaging
65437	G	T	chr13:48942736 G/T	probably damaging
61745	T	A	chr13:48939045 T/A	benign
150002	C	T	chr13:49027133 C/T	probably damaging
39561	G	A	chr13:48916850 G/A	benign
39561	G	A	chr13:48916850 G/A	benign
73828	G	A	chr13:48951129 G/A	probably damaging
156698	C	T	chr13:49033829 C/T	probably damaging
56923	T	A	chr13:48934223 T/A	probably damaging
56923	T	A	chr13:48934223 T/A	probably damaging

153353	G	C	chr13:49030485 G/C	benign
153353	G	T	chr13:49030485 G/T	benign
Potential Missense variants near position chr13:49033844 (at codon 661)				
gDNA position	Reference allele	Alternate allele	Polyphen2 format	Polyphen2 prediction
156710	C	A	chr13:49033841 C/A	benign
156711	C	A	chr13:49033842 C/A	probably damaging
156714	G	T	chr13:49033845 G/T	probably damaging
156716	C	A	chr13:49033847 C/A	probably damaging
156717	T	G	chr13:49033848 T/G	probably damaging

Table 4.7 Enrichment of missense mutation within exon 20. Comparison between observed mutations and the simulated frequency of missense changes at amino acids and codons in exon 20, to find localized pathogenicity within missense mutations. Only the significant results are reported here.

Change (Codon or Amino Acid)	99% CI of simulation	Observed variants	P value
CGG-TGG	[0, 4]	6	0
Arginine - Tryptophan	[0, 4]	6	0

Table 5.1 Predicted and observed mutations in different substitution classes. Comparison between predicted and observed mutations on a separate dataset of *de novo* germline mutations. The 95% CI of predicted mutations is reported here.

Substitution Class	95% CI of predicted mutations	Observed mutations
A-to-C	[291, 363]	341
A-to-G	[1199, 1339]	1279
A-to-T	[295, 366]	316
C-to-A	[427, 511]	430
C-to-G	[350, 430]	430
C-to-T	[1862, 2036]	1952
C-to-T at CpG only	[726, 834]	816

Table 5.2 Log likelihood of different sequence context models on a mutational data. The log likelihood of different *de novo* sequence context mutation models on a separate dataset of *de novo* germline mutations.

Sequence model	Log Likelihood
heptanucleotide	-29447
trinucleotide	-29718
1mer+CpG	-29714
1mer	-31196

Table 5.3 Log likelihood of different sequence context models on coding variant data. The log likelihood of different *de novo* coding sequence context mutation models on a separate dataset of low frequency variants from EVS dataset.

Sequence model	Log Likelihood
heptanucleotide	-3332725
trinucleotide	-3364132
1mer+CpG	-3509111
1mer	-3611400

FIGURES

Figure 2.1 Intuition behind the substitution probability model. (a) Defining the non-Bayesian probability and Bayesian posterior probability of nucleotide substitution for a 7-mer context. Here I use the example CTACGAT, where position 4 is the polymorphic site and the three nucleotides located 5′ and 3′ constitute the remainder of that site's local 7-mer sequence context. I count (i) the number of occurrences of that 7-mer context found in the reference genome and (ii) the number of times I observe a polymorphic substitution at position 4. The example shown here is a C-to-T substitution. To generate the posterior probabilities, I sum the observed counts of occurrences and substitutions with a count obtained from the modeled prior. I apply this mathematics to all 7-mer sequence contexts for all substitution classes and then merge the reverse-complementary pairs (the A-to-C class was merged with the T-to-G class, etc.). This results in a total of 24,576 parameters, each representing a unique 7-mer sequence context. (b) Illustration showing how the same 7-mer sequence context on different codon frames leads to different types of amino acid change. Depicted are three cases where a C-to-T substitution that occurs in the sequence context CTA[C/T]GAT at either position 1, 2 or 3 of a codon results in a synonymous, nonsynonymous or nonsense change in amino acid identity.

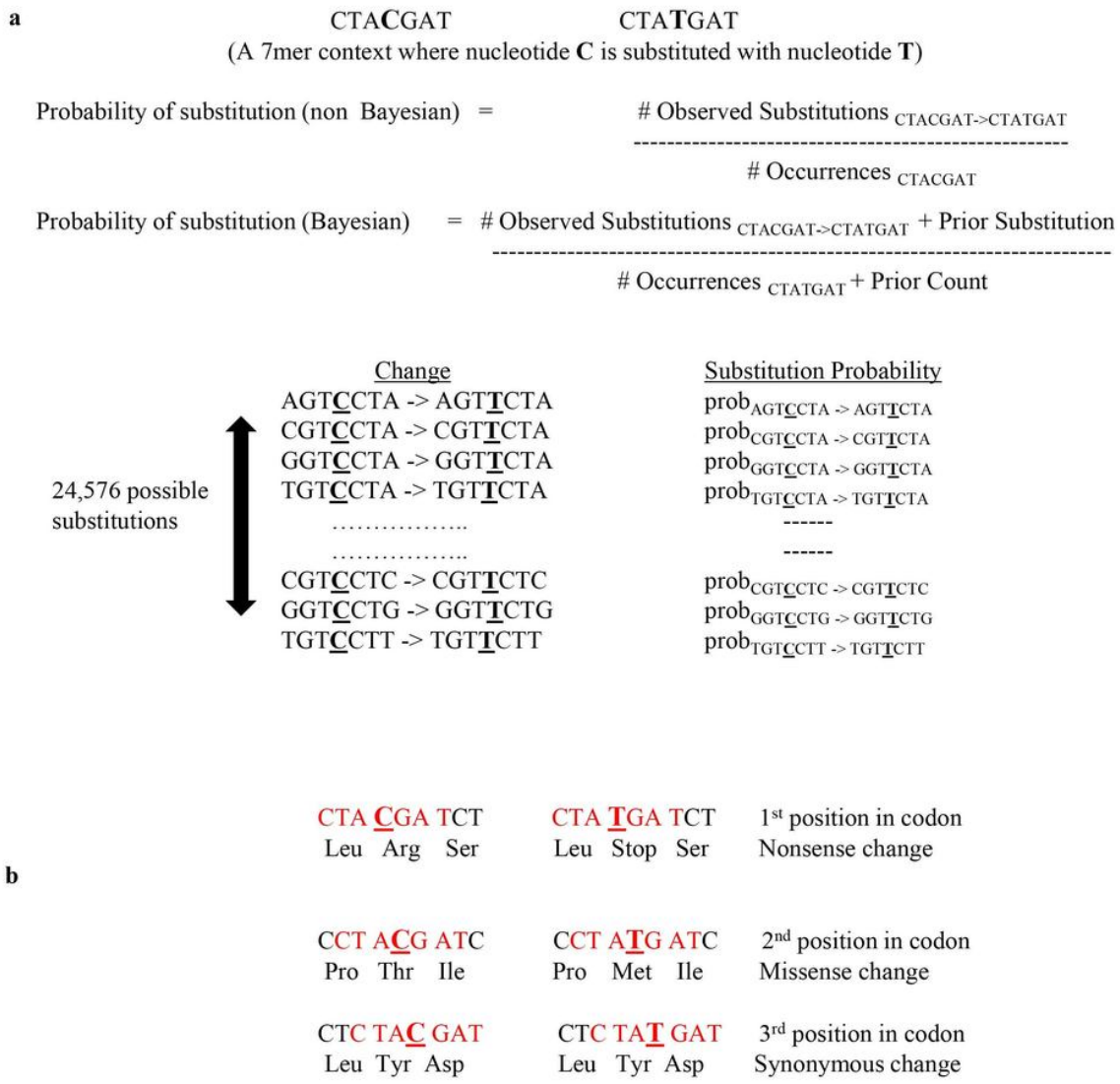


Figure 2.2 Substitution probabilities from 1KG and HapMap data. Scatter plot of nucleotide substitution probabilities inferred from 1000 genomes variants and separately from HapMap variants for each 7-mer sequence context change. The substitution probabilities in both cases are strongly correlated ($R^2 = 0.91$, $P << 10^{-100}$) with each other.

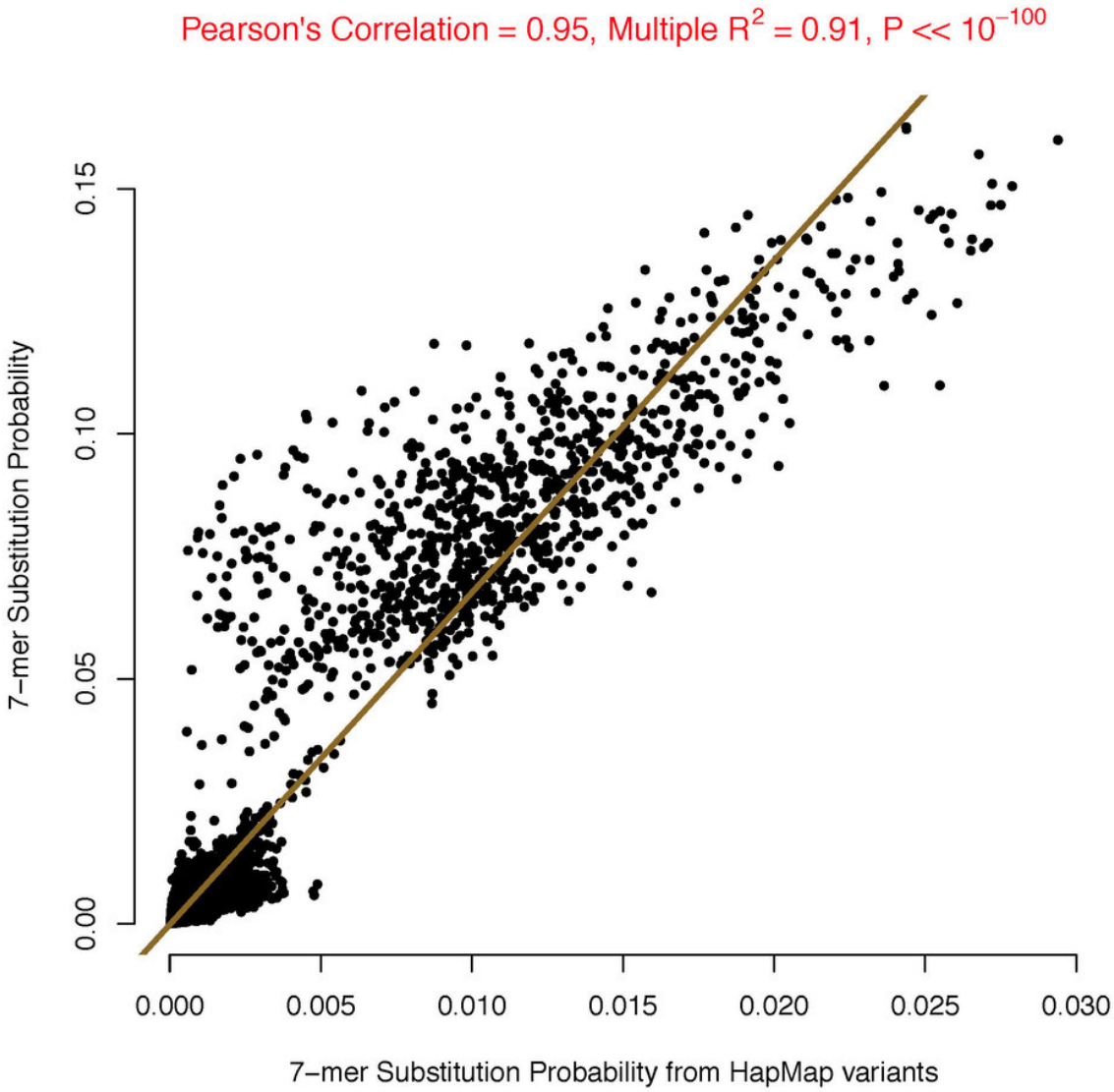


Figure 2.3 Substitution probabilities in different human populations. Genome-wide nucleotide substitution probabilities are correlated across different human populations. (a) The nucleotide substitution probabilities estimated from the 1-mer model for three human population groups (African, European, and Asian) obtained from the 1KG Project. (b) The nucleotide substitution probabilities estimated from the 7-mer context in the same three populations. Because the x-axis for this plot represents 24,576 sequence contexts, it was not practical to list them individually as was done in part A, above. The contexts are represented graphically, sorted from lowest-to-

highest nucleotide substitution probability, as observed in the African group. Data for the European and Asian groups was then represented according to the order obtained for the African group, to make comparison possible across the populations for any given sequence context.

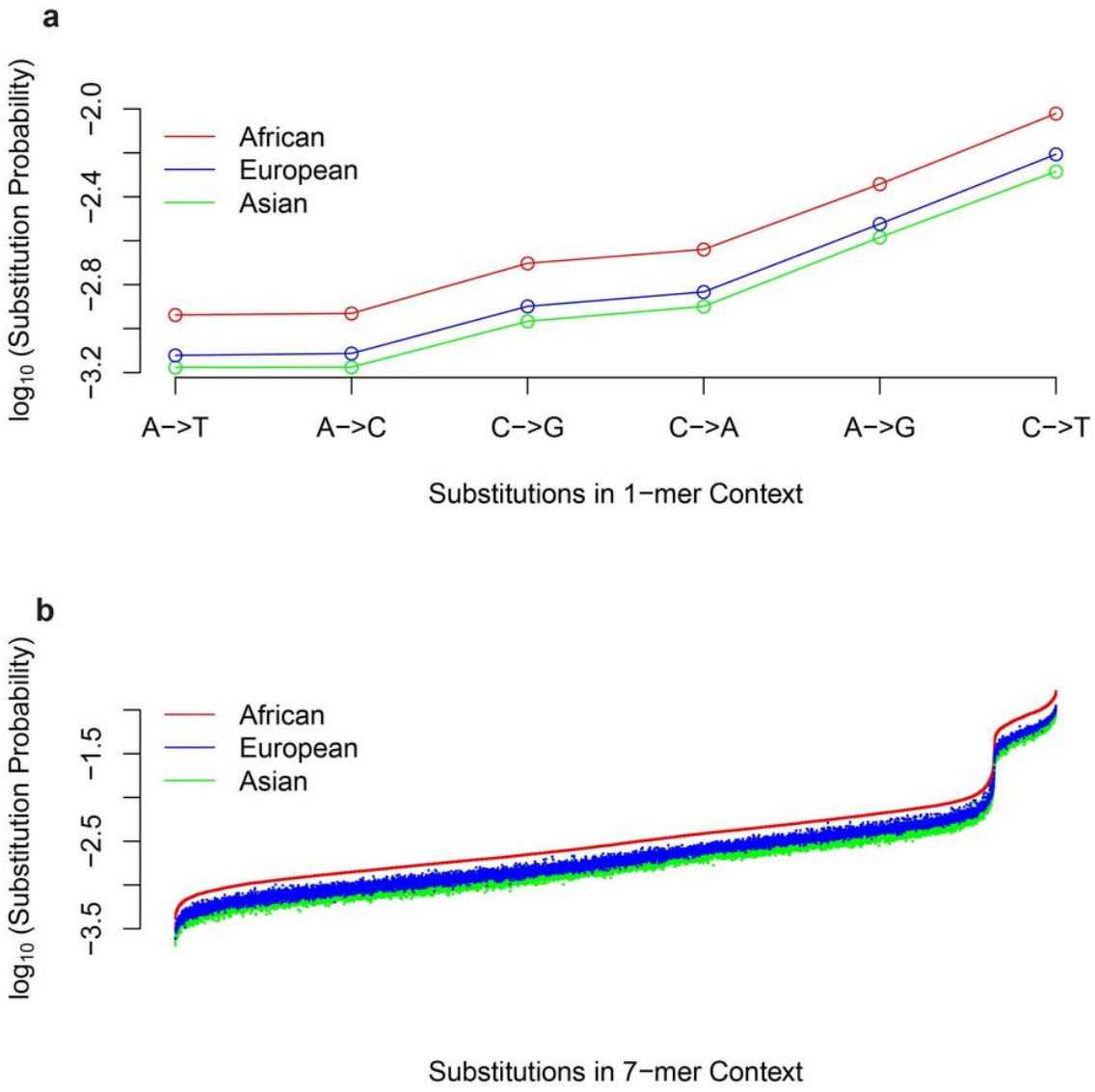


Figure 2.4 Variability in heptanucleotide substitution probabilities at CpG context. Simulations based on a fixed C-to-T substitution rate (blue) at CpG contexts do not capture the observed distribution of substitution probabilities (red) within the 7-mer sequence context. Rates predicted

from my regression model (black) closely match the substitution probabilities observed under the 7-mer sequence context (R^2 = 0.93).

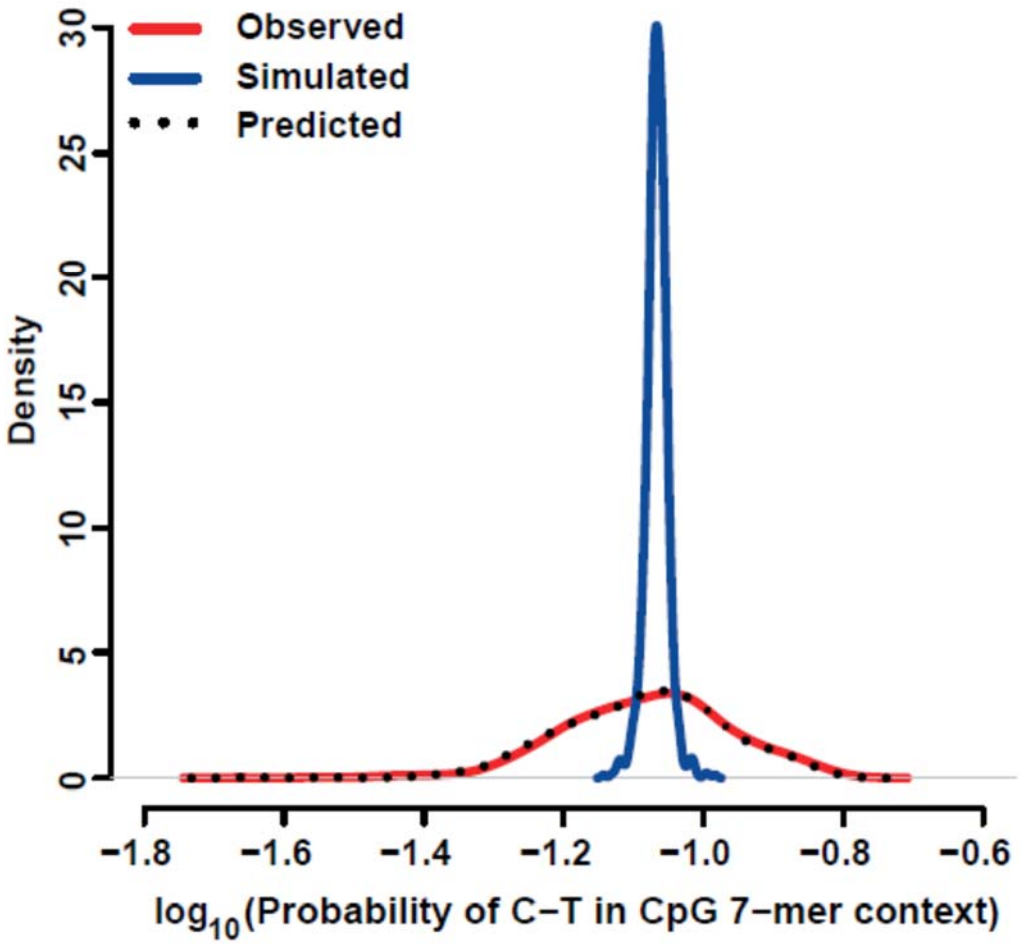


Figure 2.5 3-mer model and variability in heptanucleotide substitution probabilities at CpG context. Observed distribution of C-to-T substitution probabilities within a 7-mer CpG sequence context, compared to simulations assuming a 3-mer model. The distribution of substitution probabilities in a 7-mer context around a CpG site is significantly greater ($P << 10^{-10}$) than expected, relative to the distribution of probabilities simulated assuming a 3-mer model.

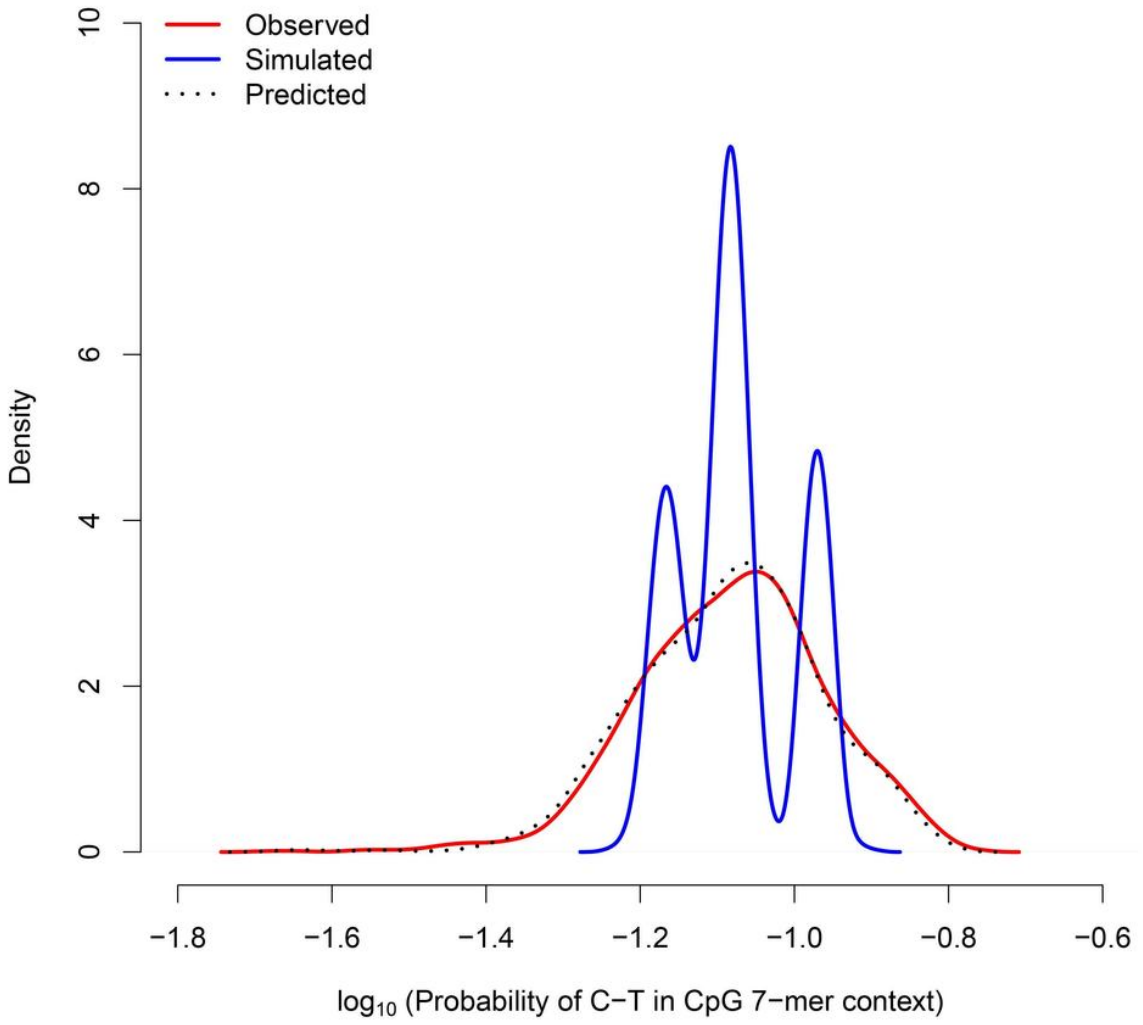


Figure 2.6 C-to-T substitution probabilities and methylation patterns. Probabilities of C-to-T substitutions for the following sequence contexts: “CpG Me-” = CpG 7-mer contexts that were unmethylated in all sperm samples[110]; “CpG Me+” = CpG 7-mer contexts that were methylated in all sperm samples. *** represents $P << 10^{-100}$.

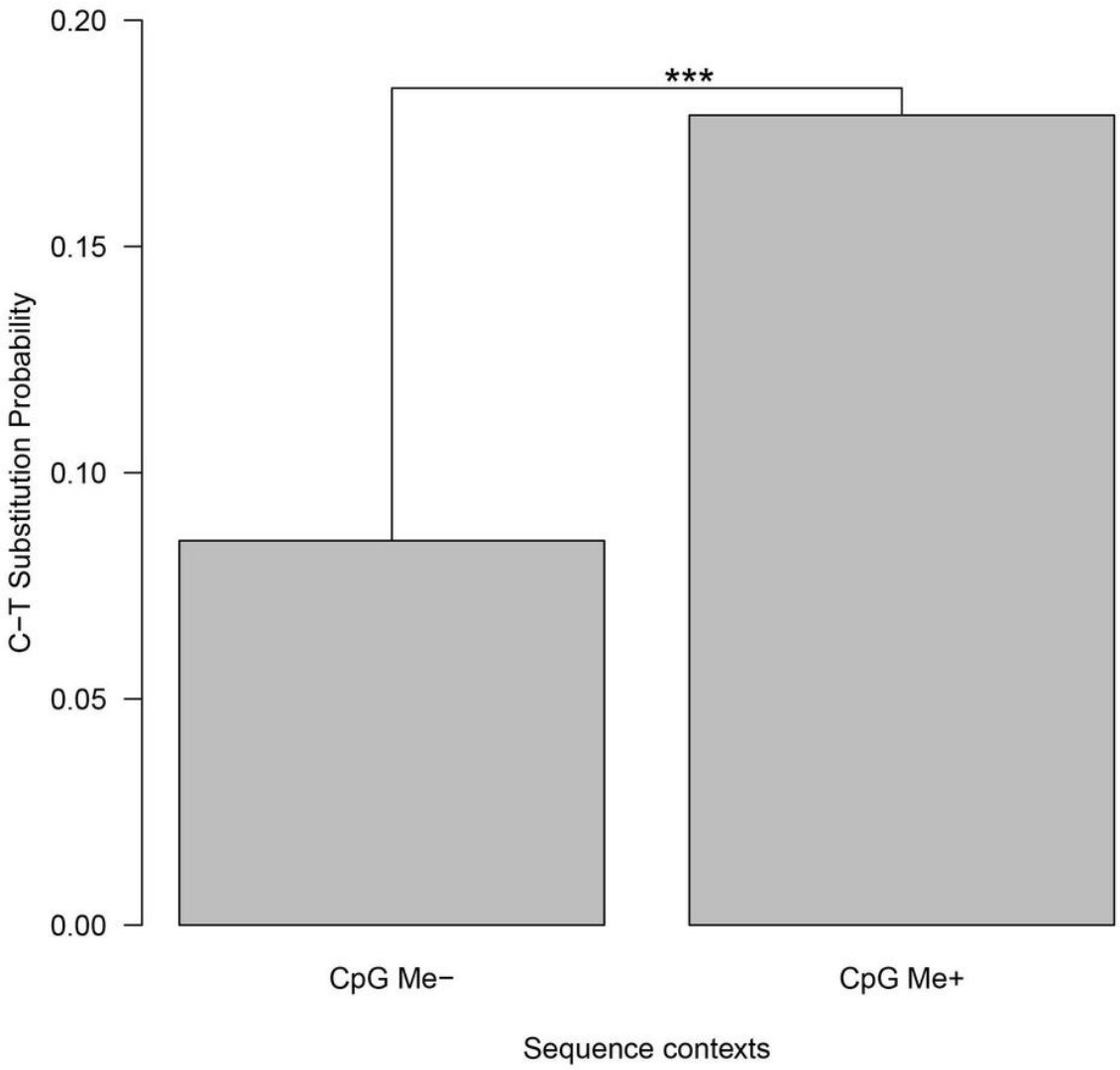


Figure 2.7 Methylation intensity and heptanucleotide substitution probabilities at CpG context. Correlation between average methylation intensity versus probability of C-to-T substitution in CpG 7-mer context.

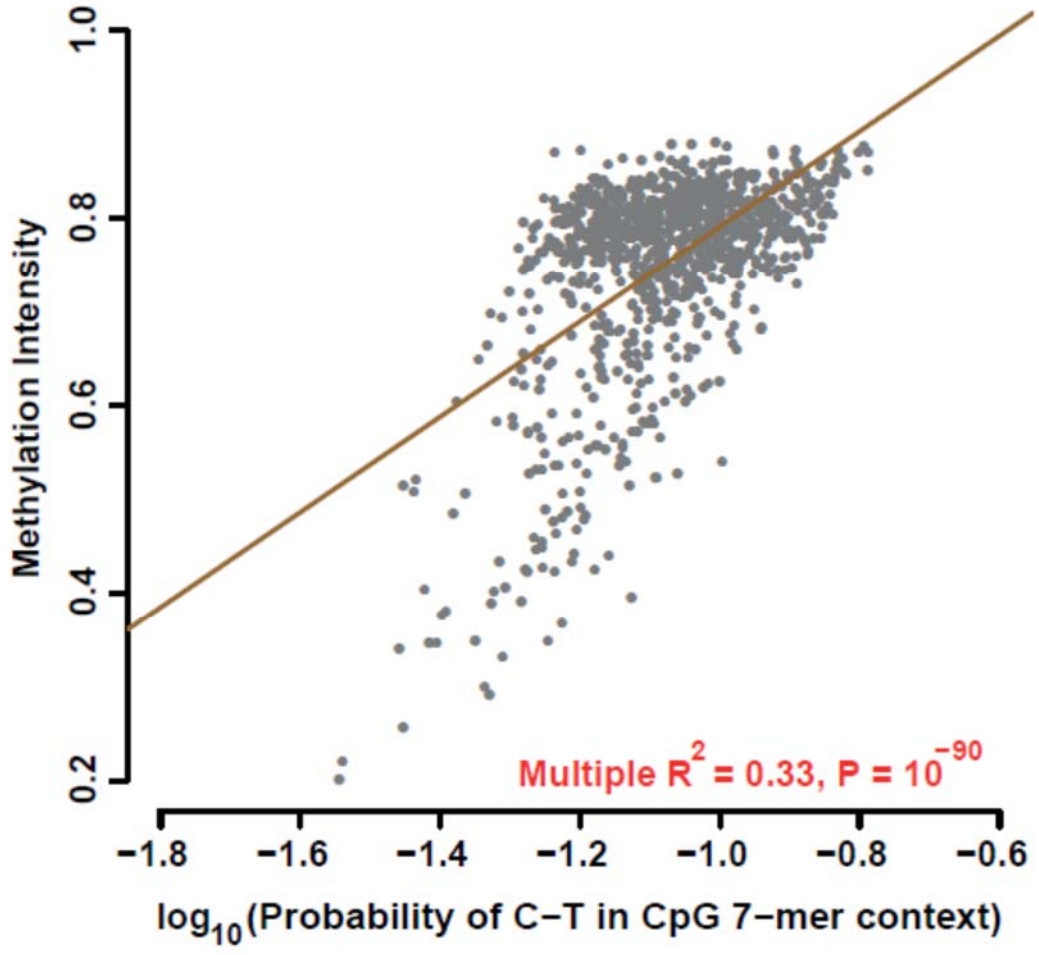


Figure 2.8 Methylation (more tissues) and substitution probabilities at CpG context. Correlation between average methylation intensity versus probability of C-to-T substitution in CpG 7-mer context. (a) Scatter plot of average methylation intensity in brain samples against substitution probability at each 7-mer CpG context. (b) Scatter plot of average methylation intensity in oocyte samples against substitution probability at each 7-mer CpG context. (c) Scatter plot of average methylation intensity in blood samples against substitution probability at each 7-mer CpG context. (d) Scatter plot of average methylation intensity in blastocyst samples against substitution probability at each 7-mer CpG context. In all cases the substitution probability is moderately correlated (R^2 ~ 0.3) with methylation intensity at each 7-mer CpG sequence context.

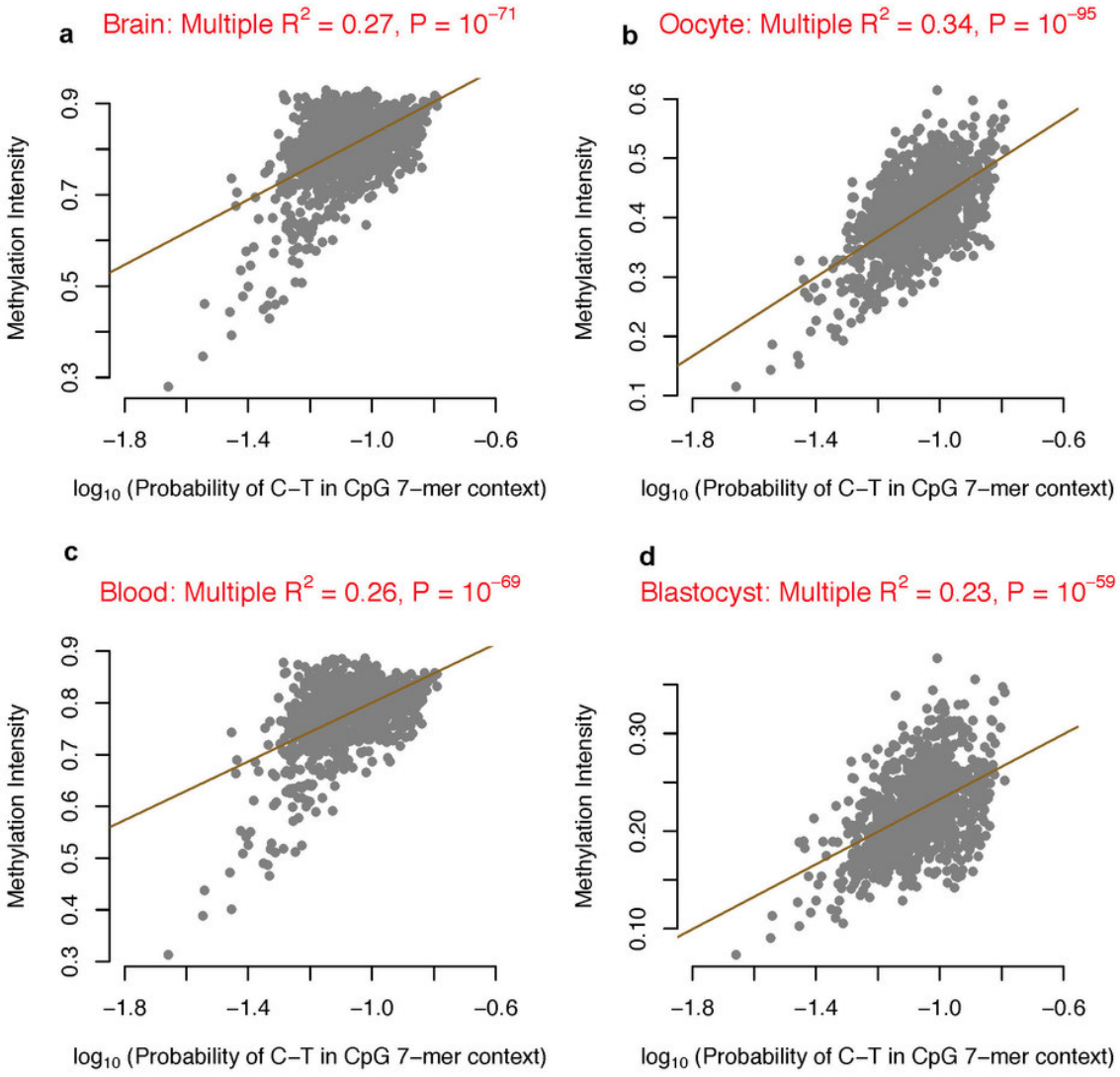


Figure 2.9 Substitution probabilities in intergenic noncoding region. Posterior probabilities of all classes of nucleotide substitution in the intergenic non-coding genome, estimated using the 7-mer context model. Sequences contexts are further stratified by color to indicate either the presence of a CpG (C at the polymorphic 4th position and G at the 5th position, for C-to-A, C-to-G and C-to-T substitution classes = CpG+; else CpG-) or the ApT state (A at the polymorphic 4th position and T at the 5th position, for A-to-G and A-to-T substitution classes = ApT+; else ApT-). For A-to-

C, the ApT state did not significantly contribute to variability in the estimated probability distribution.

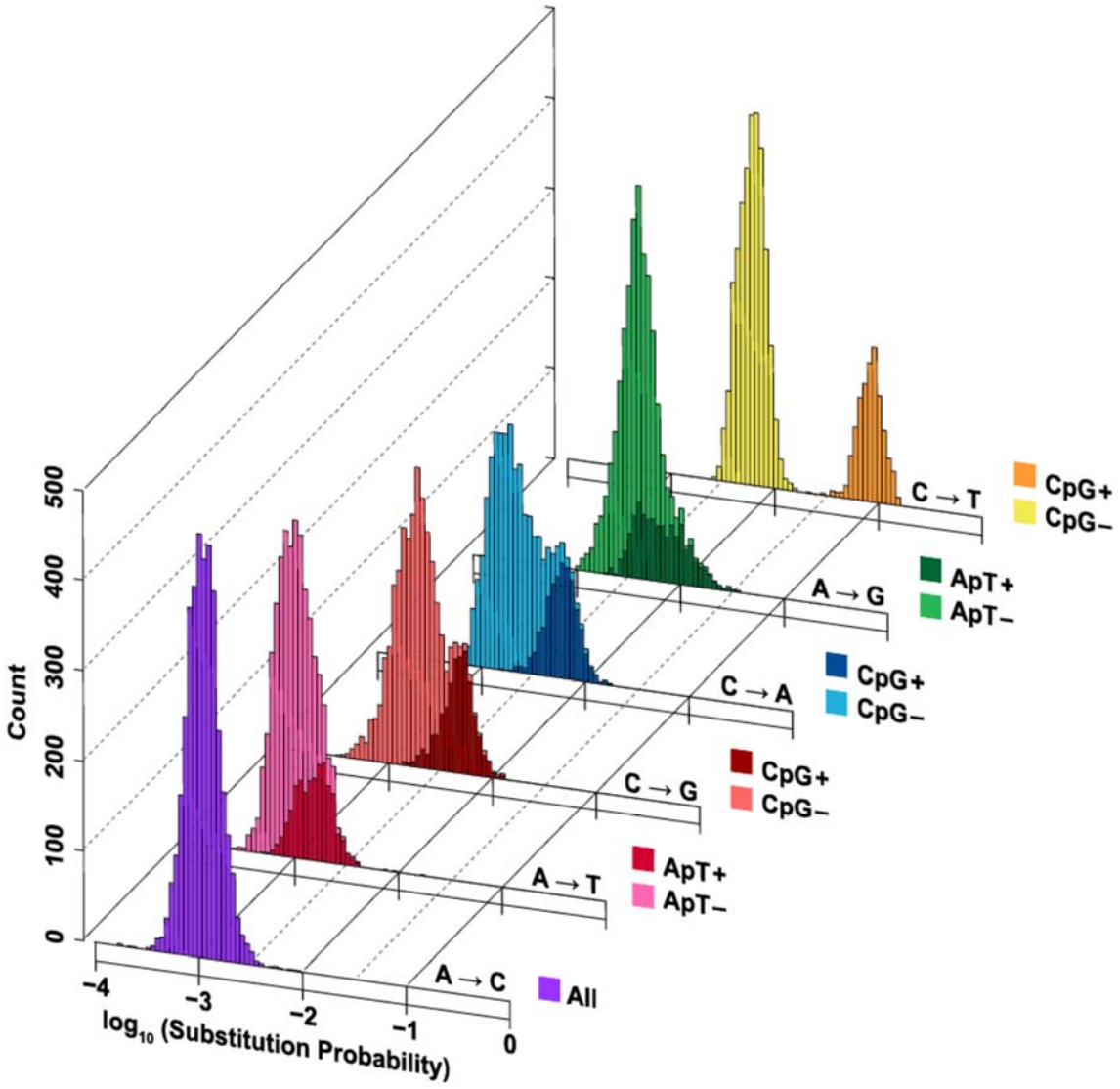


Figure 2.10 Distance from nearest gene and CpG sites. Box-and-Whisker plot of the distances between sequence contexts that contains a CpG site (C at polymorphic 4th position, fixed G at 5th position) and the gene nearest to that context found in the human reference genome. “LOW” plots the distances from sequences contexts identified in the bottom 1% smallest substitution probabilities in the C-to-T substitution class (n=10). “ALL” represents the distances from all

sequences contexts containing a CpG (n=1024). “HIGH” represents the distances from sequences contexts in the top 1% smallest substitution probabilities from C-to-T substitution class (n=10). Each distribution is significantly different from one other (pair-wise each are $P << 10^{-100}$ by Wilcoxon Rank-Sum Test).

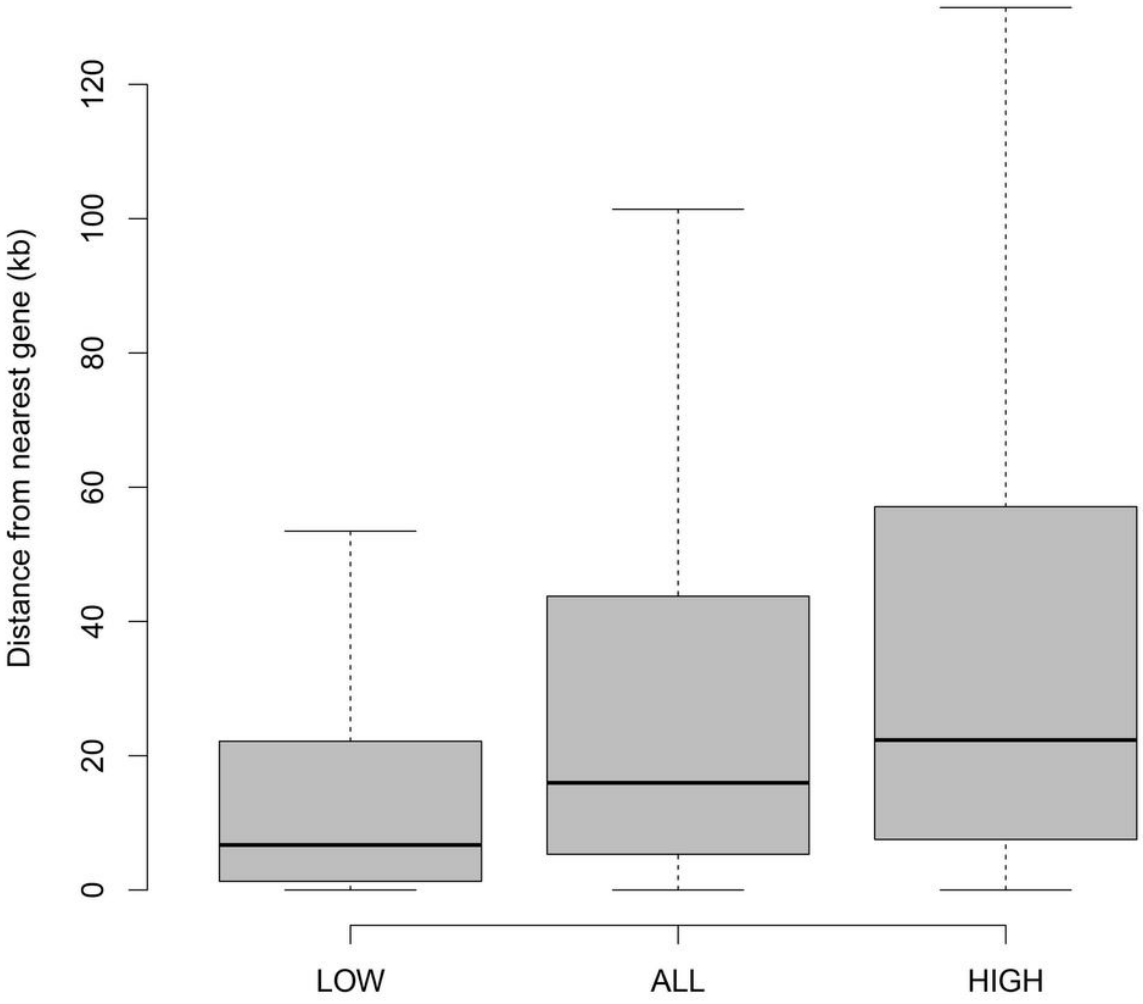


Figure 2.11 Methylation intensity at CpG sites. Box-and-Whisker plot of methylation intensity values in various sequence contexts containing a CpG site. Methylation intensity represents the average intensity values across all sperm samples[110]. “Poly-CpG” represents sequence contexts which segregate additional CpG dinucleotides beyond the CpG site at position 4 and 5 (note that a 7-mer sequence context with a CpG site can segregate up to 2 additional CpG

dinucleotides). Each distribution is significantly different from one other (pairwise each are $P < 10^{-5}$ by Wilcoxon Rank-Sum Test).

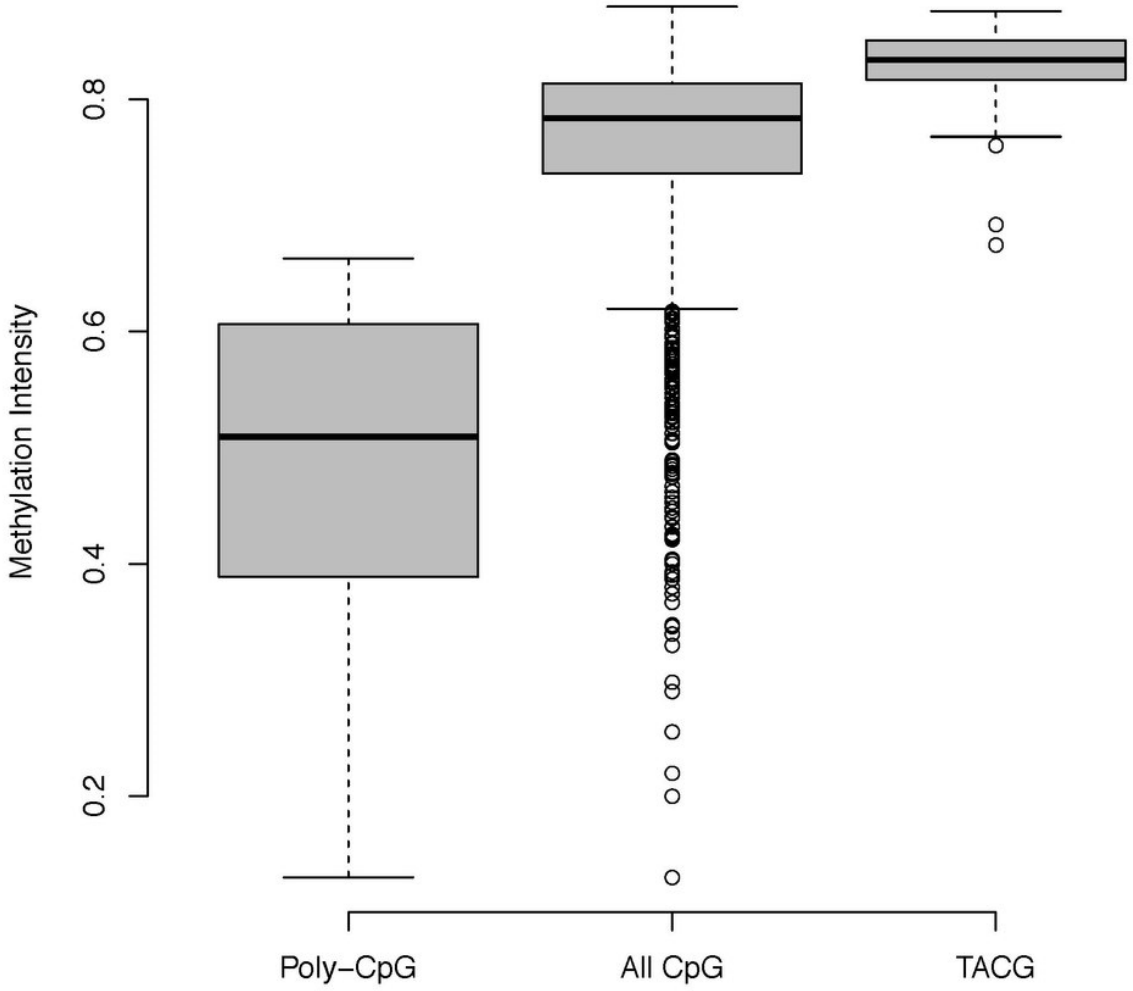


Figure 2.12 Substitution probabilities and recombination rate. Scatter plot of nucleotide substitution probabilities inferred from only 1000 genomes high recombination (rate > 3 cM/Mb in YRI population) and separately for low recombination (rate < 0.05 cM/Mb in YRI population) regions for each 7-mer sequence context change. The substitution probabilities in both cases are strongly correlated ($R^2 = 0.97$, $P << 10^{-100}$) with each other.

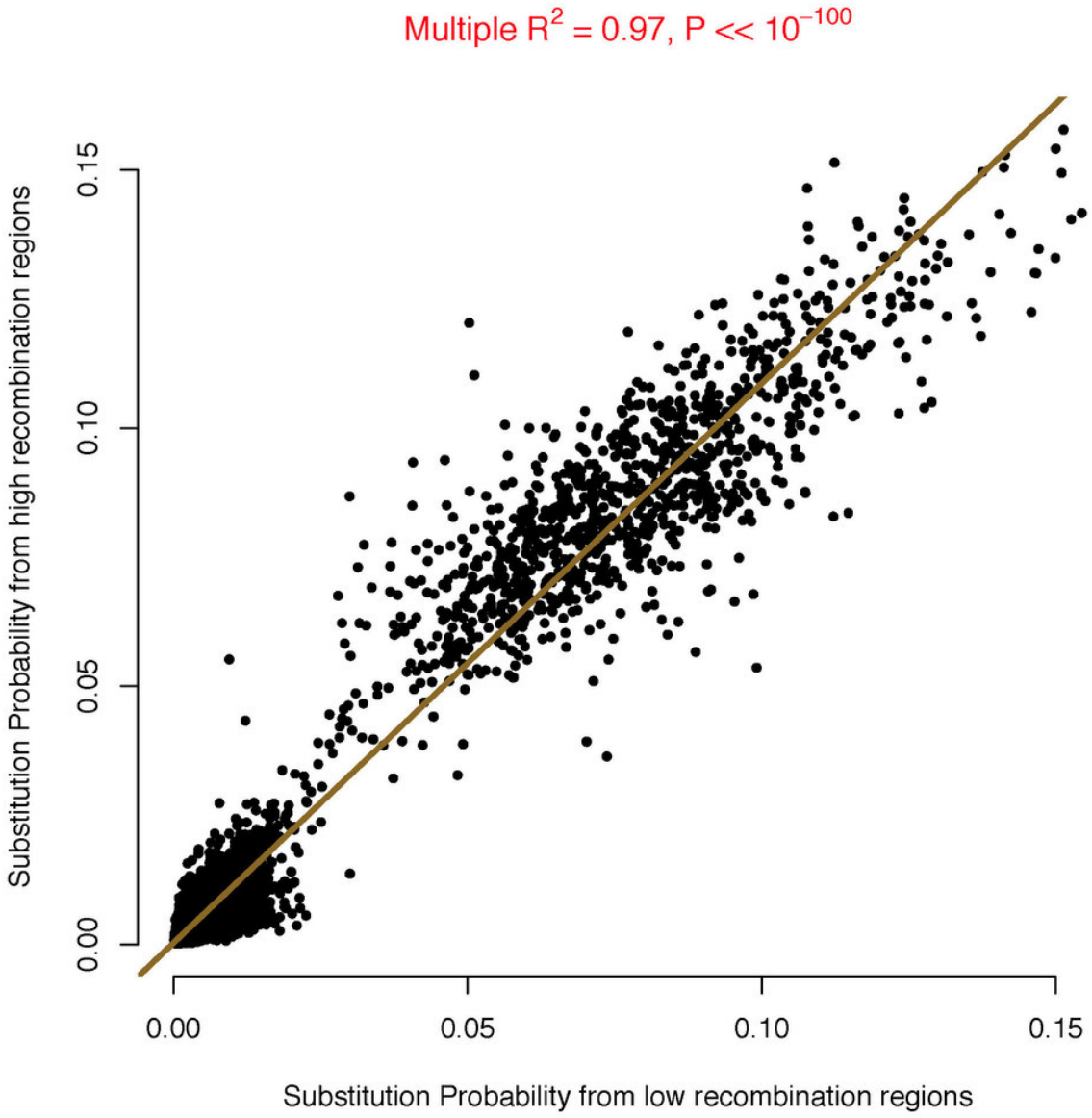


Figure 2.13 Substitution probabilities and human primate divergence. Human substitution probabilities are strongly correlated with Human-Chimpanzee and Human-Macaque divergence rates. (a) Scatter plot of nucleotide substitution probabilities against nucleotide divergence rates between human-chimpanzee at each 7-mer sequence context. (b) Scatter plot of nucleotide substitution probabilities against nucleotide divergence rates between human-macaque at each

7-mer sequence context. In both cases the substitution probabilities and divergence rates are strongly correlated (R^2 = 0.96, P << 10^{-100}) with each other.

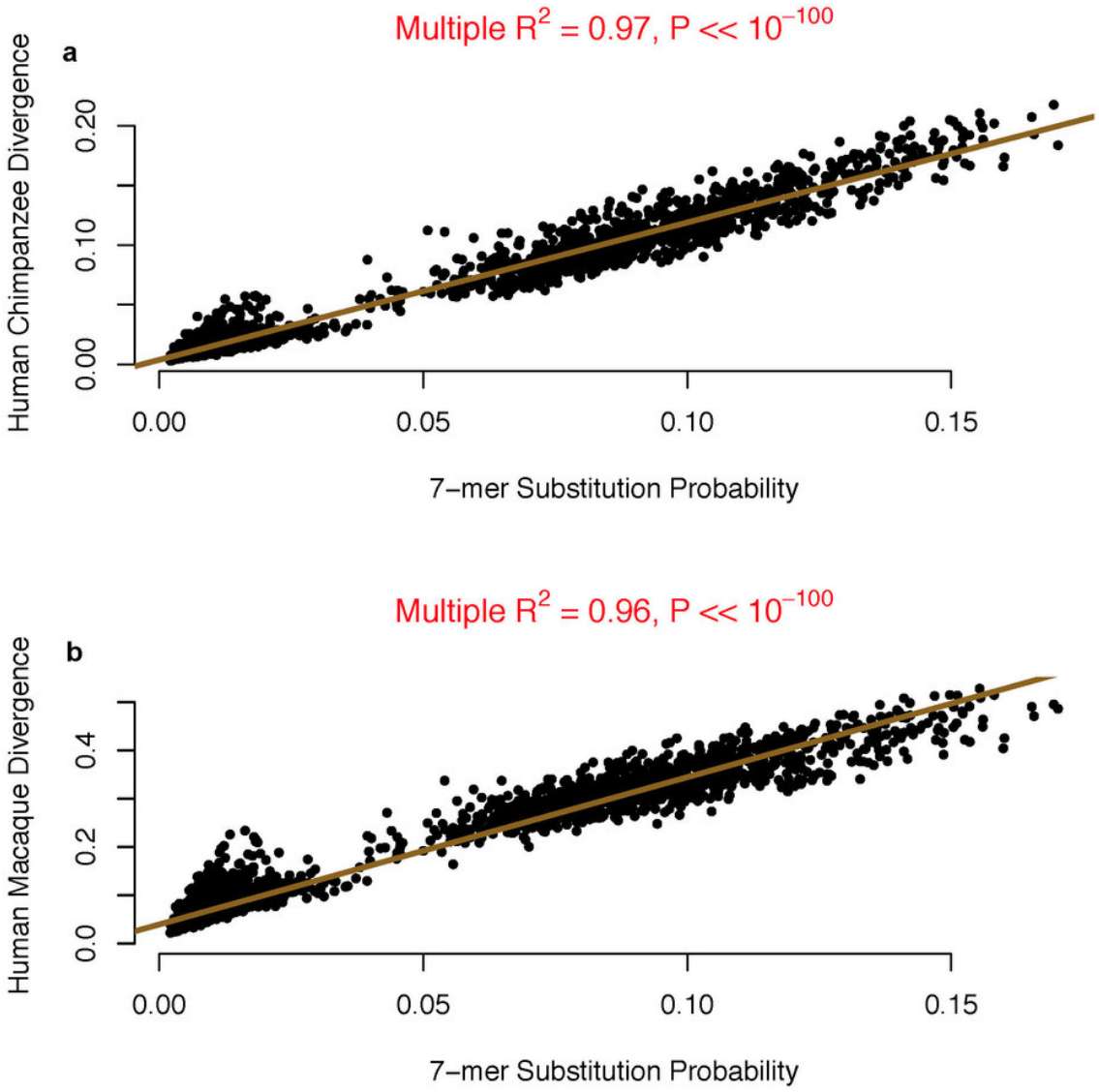


Figure 2.14 Substitution probabilities across variant frequency spectrum. Scatter plot of nucleotide substitution probabilities inferred from only 1000 genomes low frequency (1% and above MAF) and separately from rare (singletons and doubletons only) variants for each 7-mer

sequence context change. The substitution probabilities in both cases are strongly correlated (R^2 = 0.98, P << 10^{-100}) with each other.

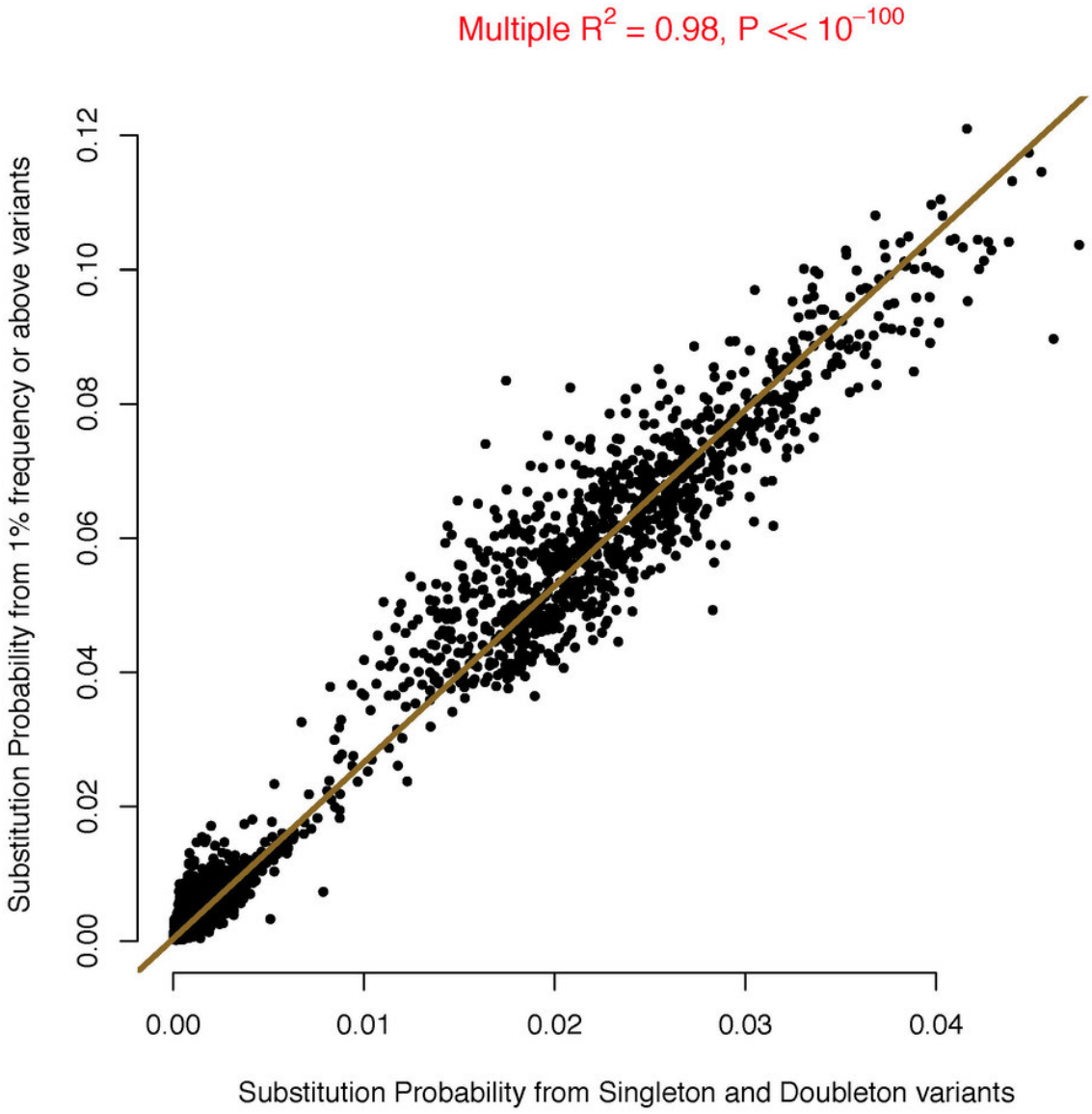


Figure 3.1 Substitution probabilities in the coding region. Posterior probabilities of nucleotide substitution for each type of amino acid substitution in the coding genome, estimated using the 7-mer coding context model. Sequences contexts are further stratified by color to indicate either the presence of a CpG (C at the polymorphic 4th position and G at the 5th position, for C-to-A, C-to-G

and C-to-T substitution classes = CpG +; else CpG -), and where evidence of substitution was only observed in the intergenic region. The presented inset zooms in specifically on the distribution for nonsense substitutions.

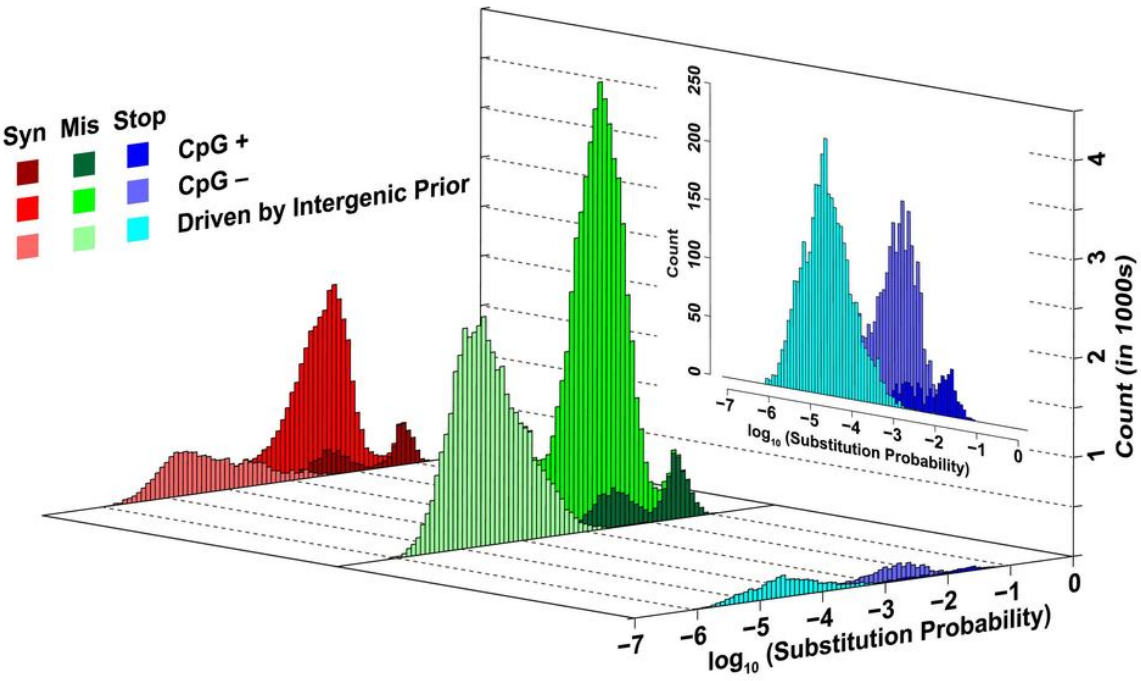


Figure 3.2 Variability in coding substitution probabilities. Violin plot for trends in amino acid replacement types across different amino acids. In (a), note the mean probability is different between Glycine and Tyrosine substitutions, though the expected trend holds (synonymous > missense > nonsense). In (b), some amino acids substitutions deviate from this expected trend, owing to the CpG context in the coding genome.

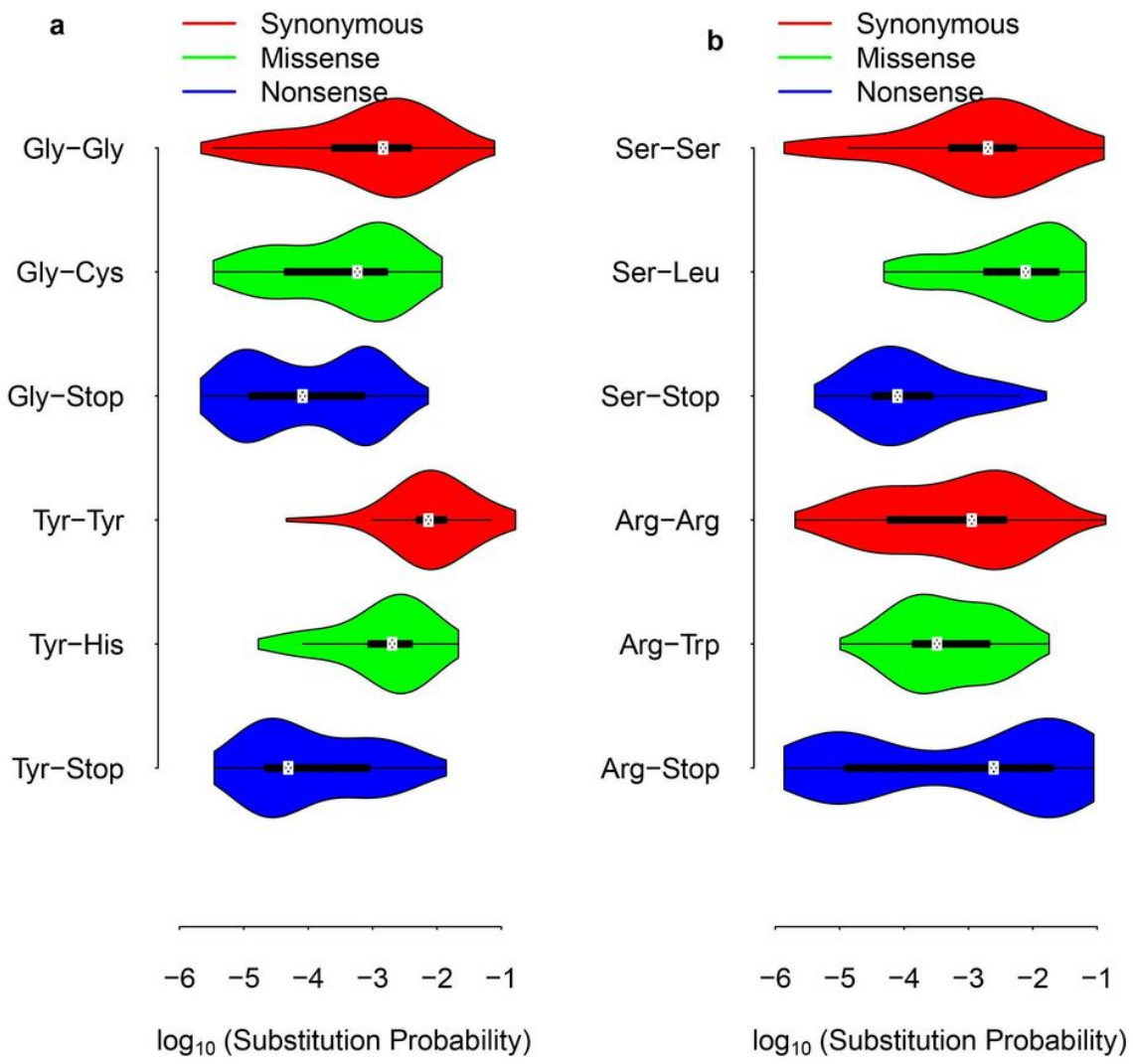


Figure 3.3 Prioritizing pathogenic variants using heptanucleotide substitution probabilities. $\log_{10}$ ratios of substitution probabilities from the 7-mer sequence context model using coding sequences matched to the intergenic non-coding sequences, for each type of substitution for all variants in the 1KG project or Human Gene Mutation Database (HGMD). Larger values indicate fewer substitutions in the coding genome than expected from matched non-coding sequences, consistent with the action of selective constraint. *** represents $P << 10^{-100}$ and ** represents $P < 10^{-29}$.

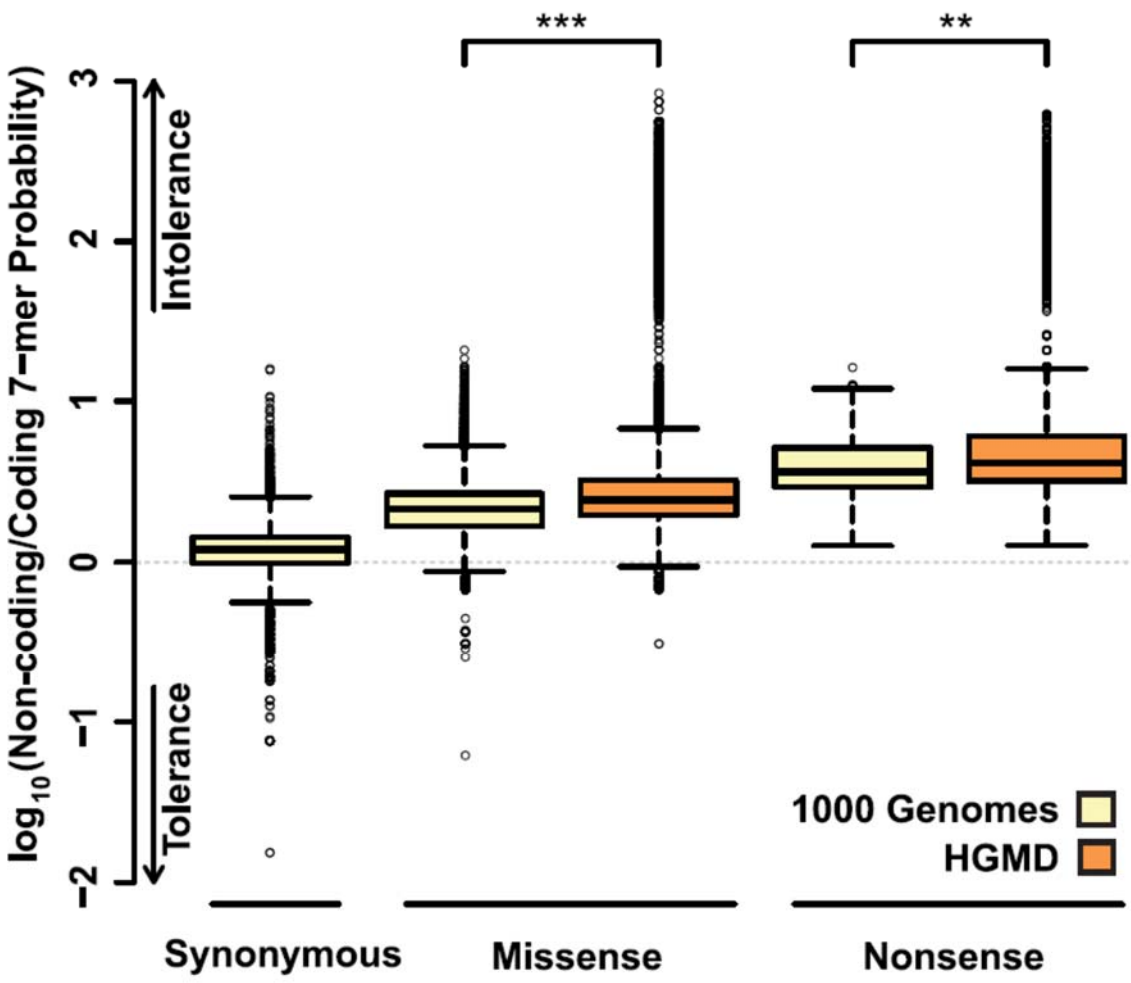


Figure 3.4 Prioritizing pathogenic variants using trinucleotide substitution probabilities. $\log_{10}$ ratios of substitution probabilities for the 3-mer model with codon context for coding sequences matched to non-coding sequences for each type of amino acid replacement. I consider all variants from the 1KG project (African, yellow) or the Human Gene Mutation Database (HGMD, orange). Larger values indicate fewer substitutions in the coding genome than expected from matched non-coding sequences (intolerance), consistent with selective constraint acting on these replacements. ** indicates $P < 10^{-53}$ and NS indicates not significant by Wilcoxon Rank-Sum Test.

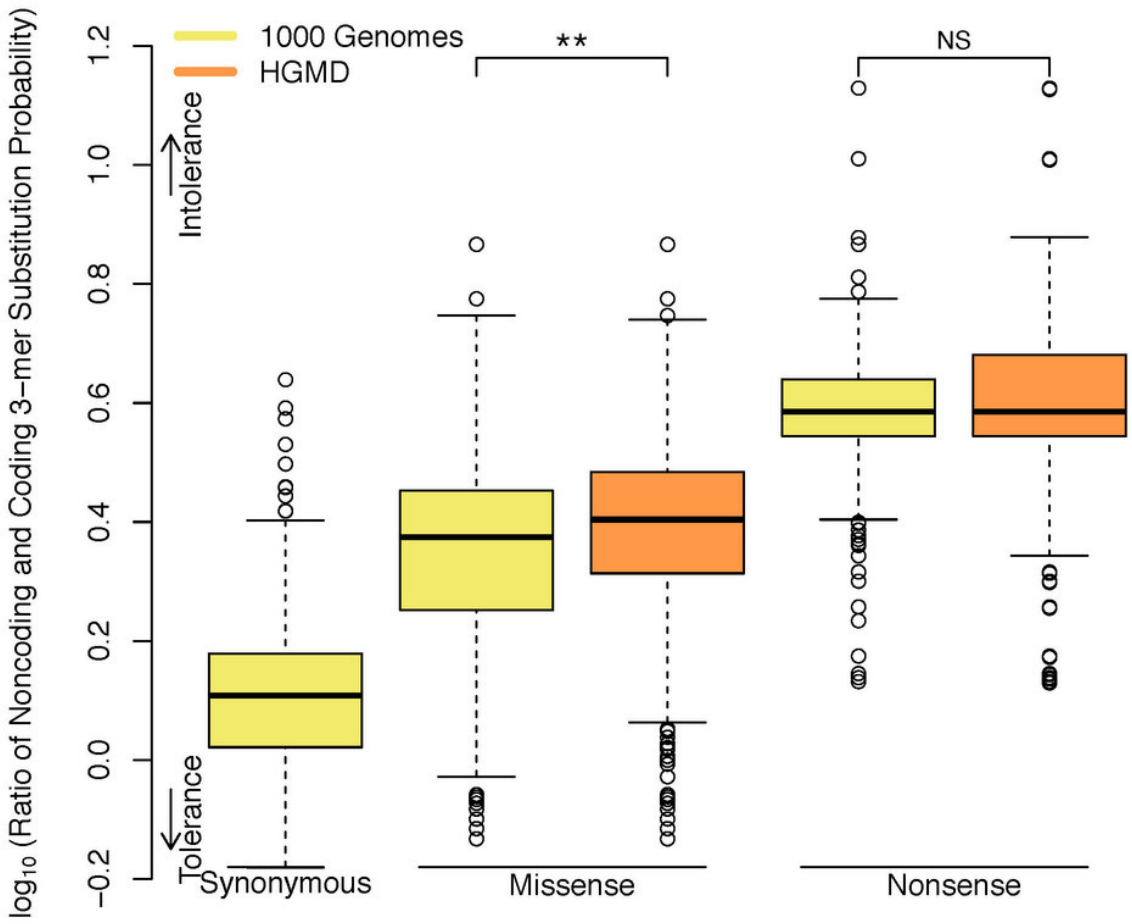


Figure 3.5 Gene scores from 1KG and EVS datasets. Gene scores based on the 7-mer coding context model calculated in 1KG (European) or EVS (European) datasets.

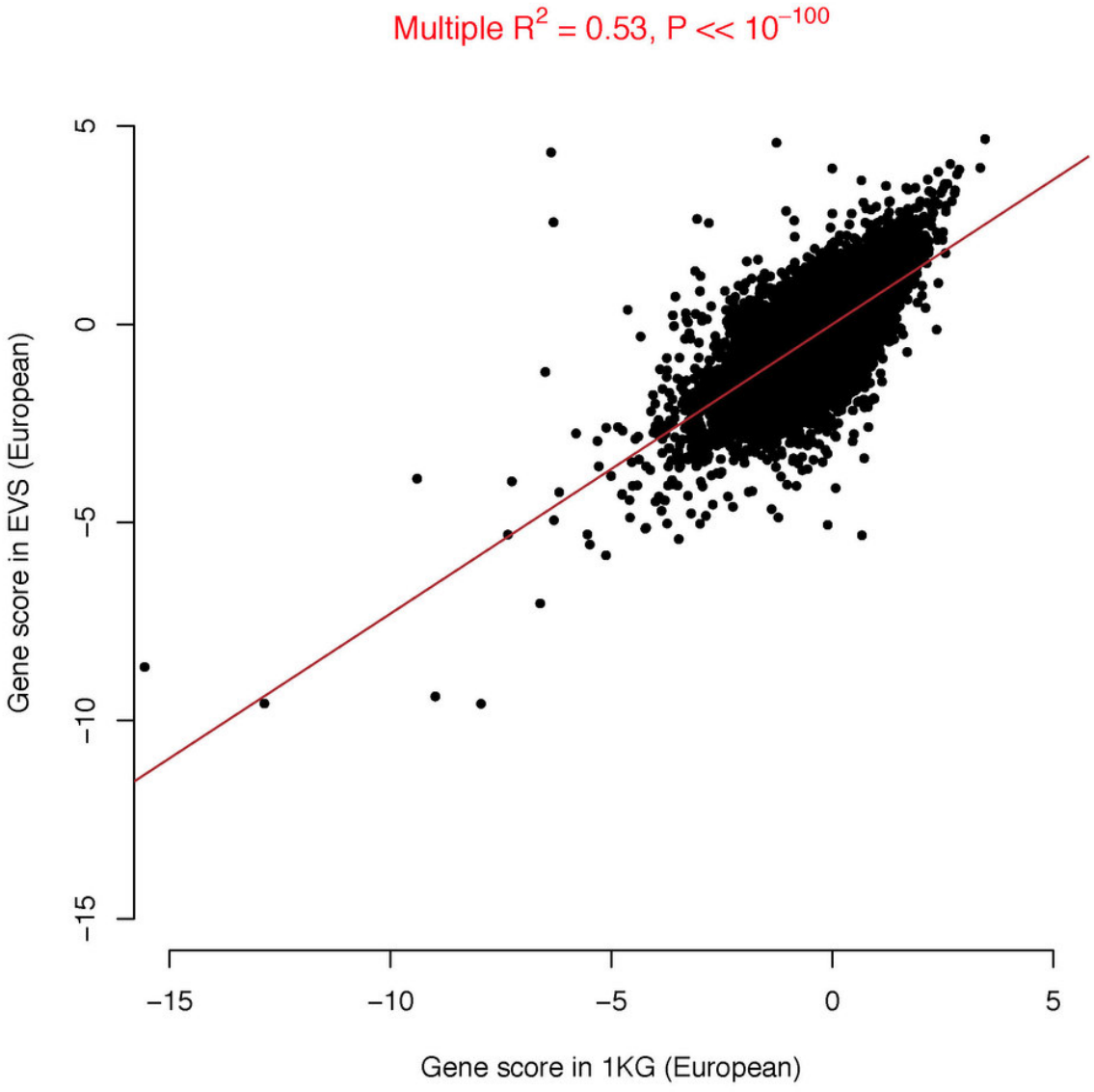


Figure 3.6 Gene scores across different gene sets. Box and whisker plot of gene scores from the model, stratified into statistically significant gene classes. Positive gene scores indicate intolerance to substitutions that change an amino acid.

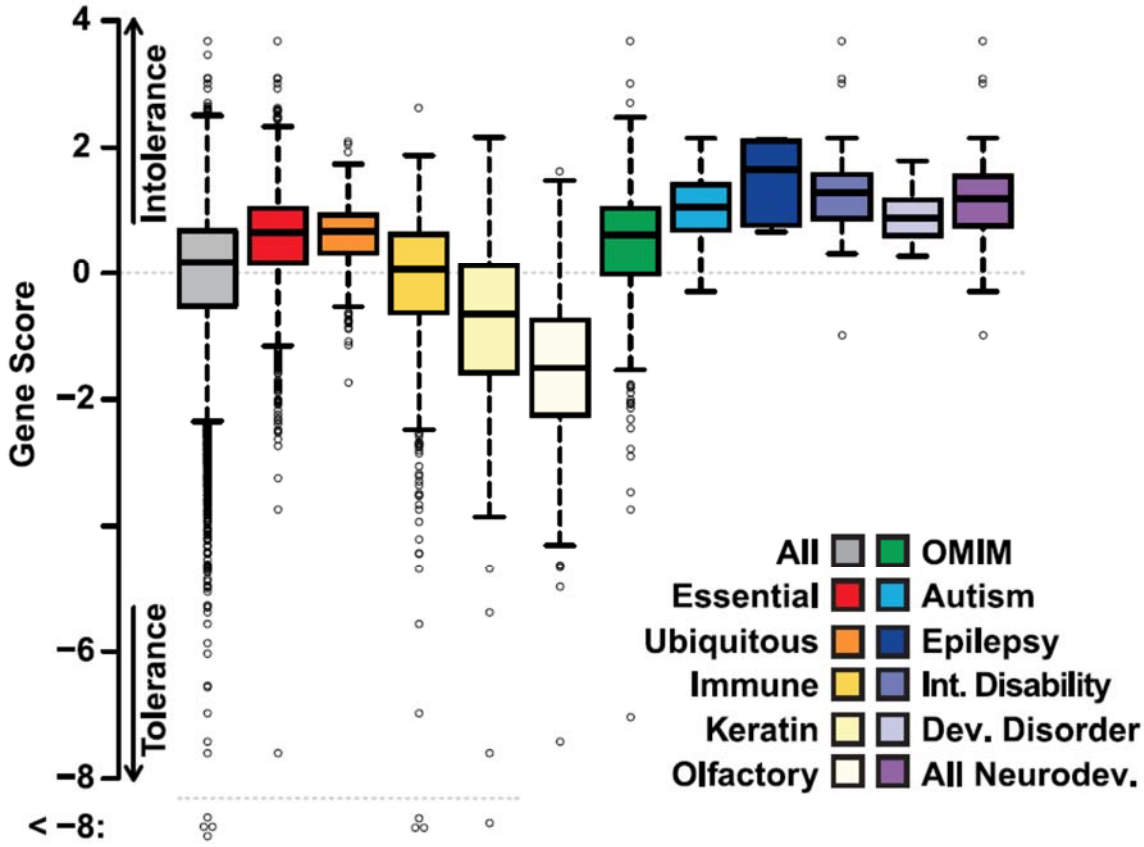


Figure 3.7 Comparison of various gene score measures. Comparison of my presented gene score (Aggarwala) built from the 1KG African group using the coding 7-mer model to scores presented by (a)(Petrovski et. al.) or (b)(Samocha et. al.). Note that in (A) all HGNC genes ids could not be mapped to Ensembl 75 genes and in (B), only a subset of gene scores was publicly available.

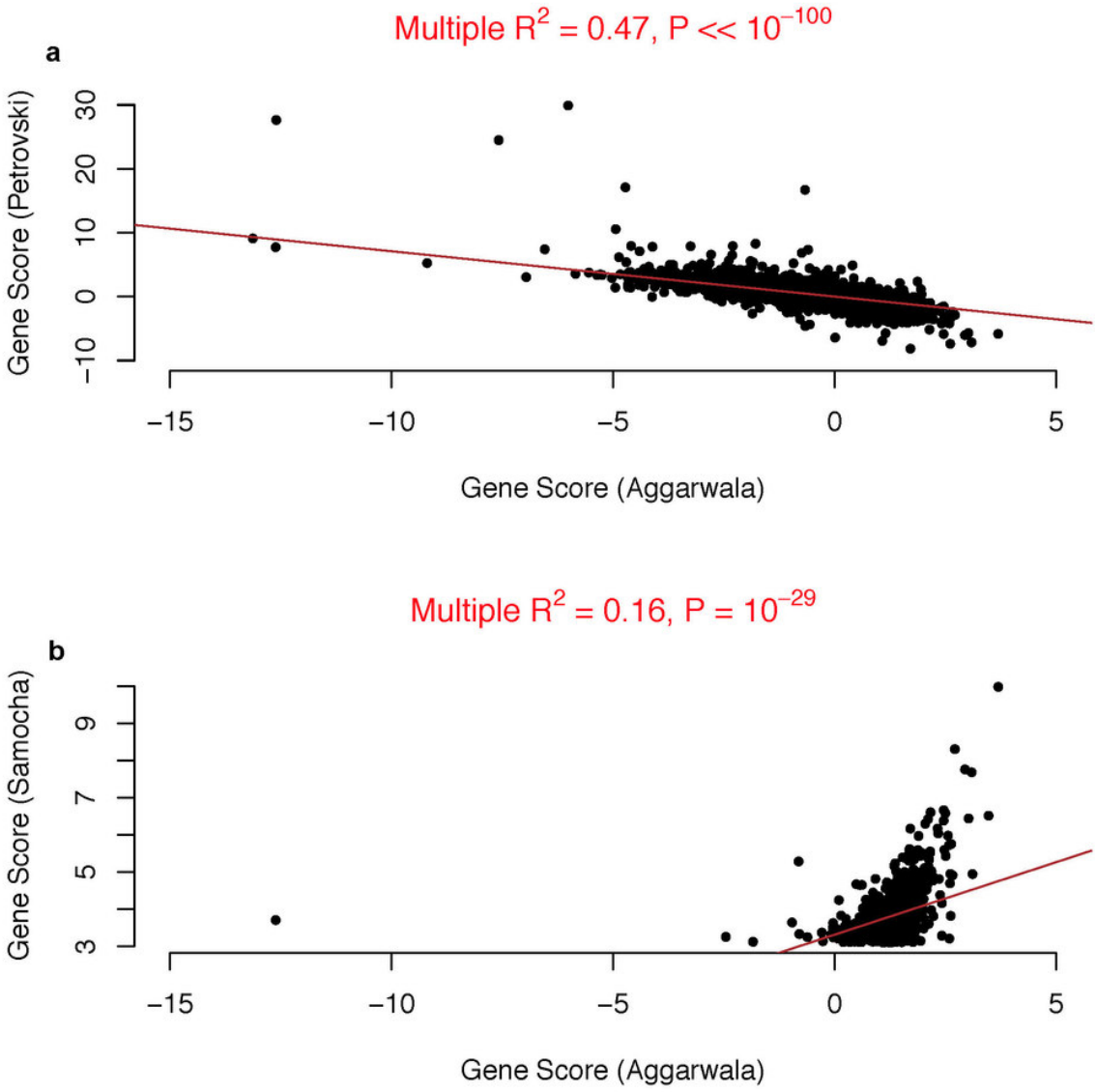


Figure 3.8 Gene scores and Autism de novo mutations. Forest plot of Odds Ratio (OR) and the 95% confidence interval (CI), and P-value when comparing the *de novo* mutational burden in cases versus controls, on intolerant genes (including or excluding known Autism genes) using different gene scoring methods. Aggarwala indicates gene score from this report, while Samocha and Petrovski refers to the intolerant gene list from those works respectively.

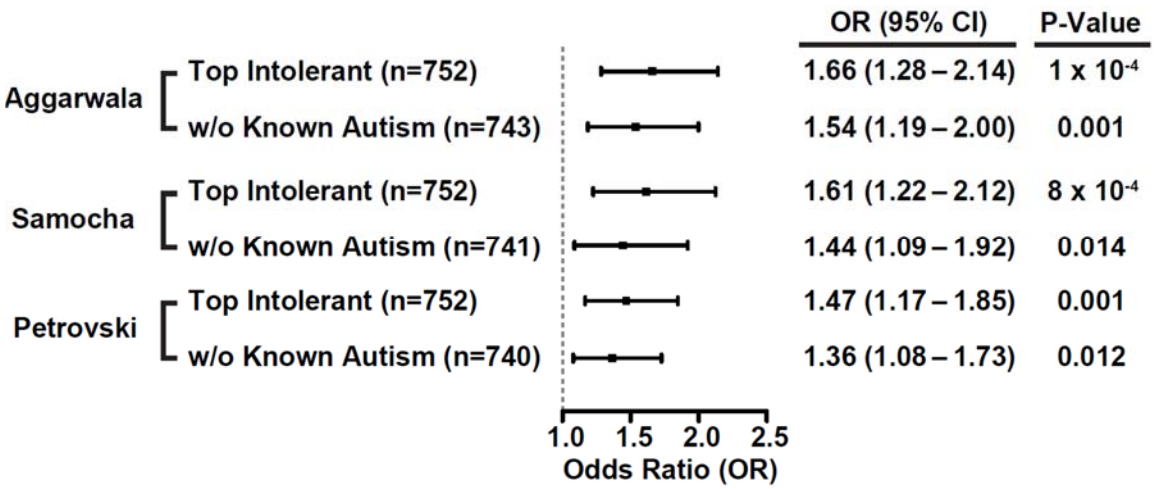


Figure 3.9 Amino acid scores and Autism de novo mutations. Forest plots of the mean amino acid score (95% CI) found from *de novo* mutations from various gene collections. Average score was based on variants ascertained in cases, except where noted (i.e., the first row: All genes in Controls). W/o: without. AC: indicates allele count for missense or nonsense changes only. * represents $P < 0.01$.

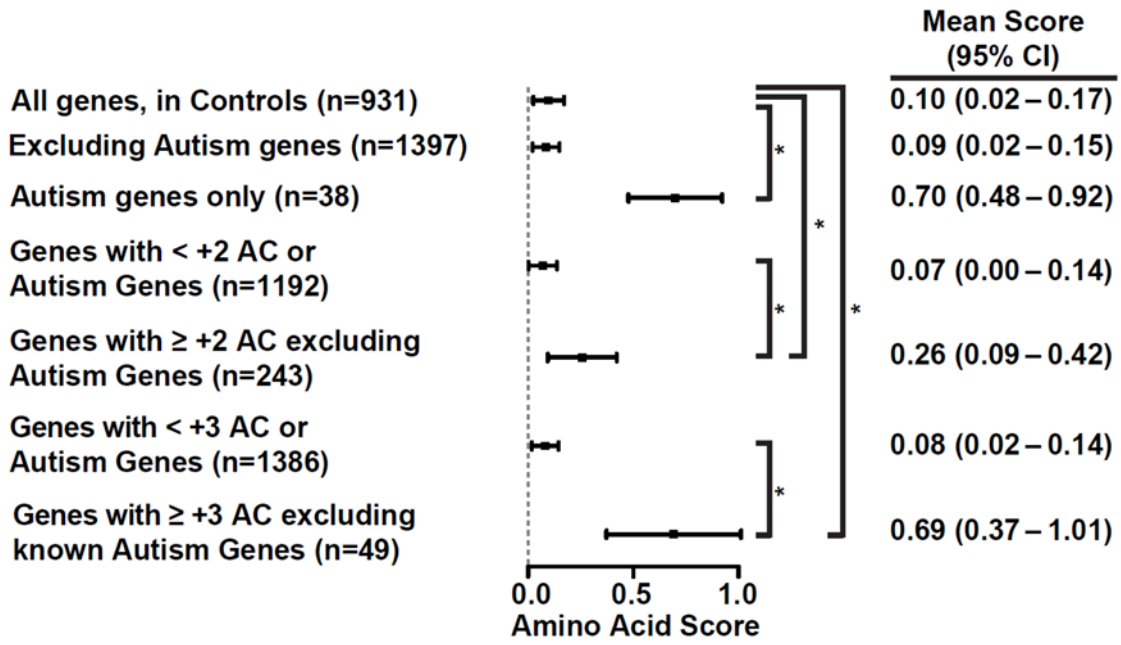


Figure 4.1 Algorithm to quantify unusual patterns of de novo mutations within a class. My approach involves three steps. First, I identify the genomic target (base pair territory) in which mutations will be characterized, and the total number of mutations found in that territory. I then distribute this total number of mutations over the target territory using a background model of mutation rate. Second, I find the expected number of mutations in different categories (Exon, mutational type like Nonsense or specific Amino Acid) using the previous distribution samples. Third and finally, I compare this to the observed number of mutation to detect statistical enrichment in a category beyond expectation. In this toy example depicted here, I focus on the genomic territory that can generate nonsense mutation (shown in red), and imagine that I have identified 10 *de novo* mutations that are nonsense. First, I identify eligible base pairs and that can result in a nonsense change. Next, I calculate the probability of mutation at each eligible base pair as the sum of substitution probabilities of that sequence context changing to a stop codon (shown in red). Second, I then distribute the mutations over multiple simulations from a multinomial distribution, and find the distribution of the expected number of mutations at each of these eligible base pairs. I am particularly interested in cases where the observed number of

mutations at a subclass (exon or an amino acid) is greater than what I see in simulations, as this is compatible with disease-relevant pathogenicity for this class of mutation, or position where the mutation(s) is located. Third and finally, for a particular subclass I combine the expected mutations at different eligible base pairs and compare the overall expected distribution with observed, and conclude enrichment.

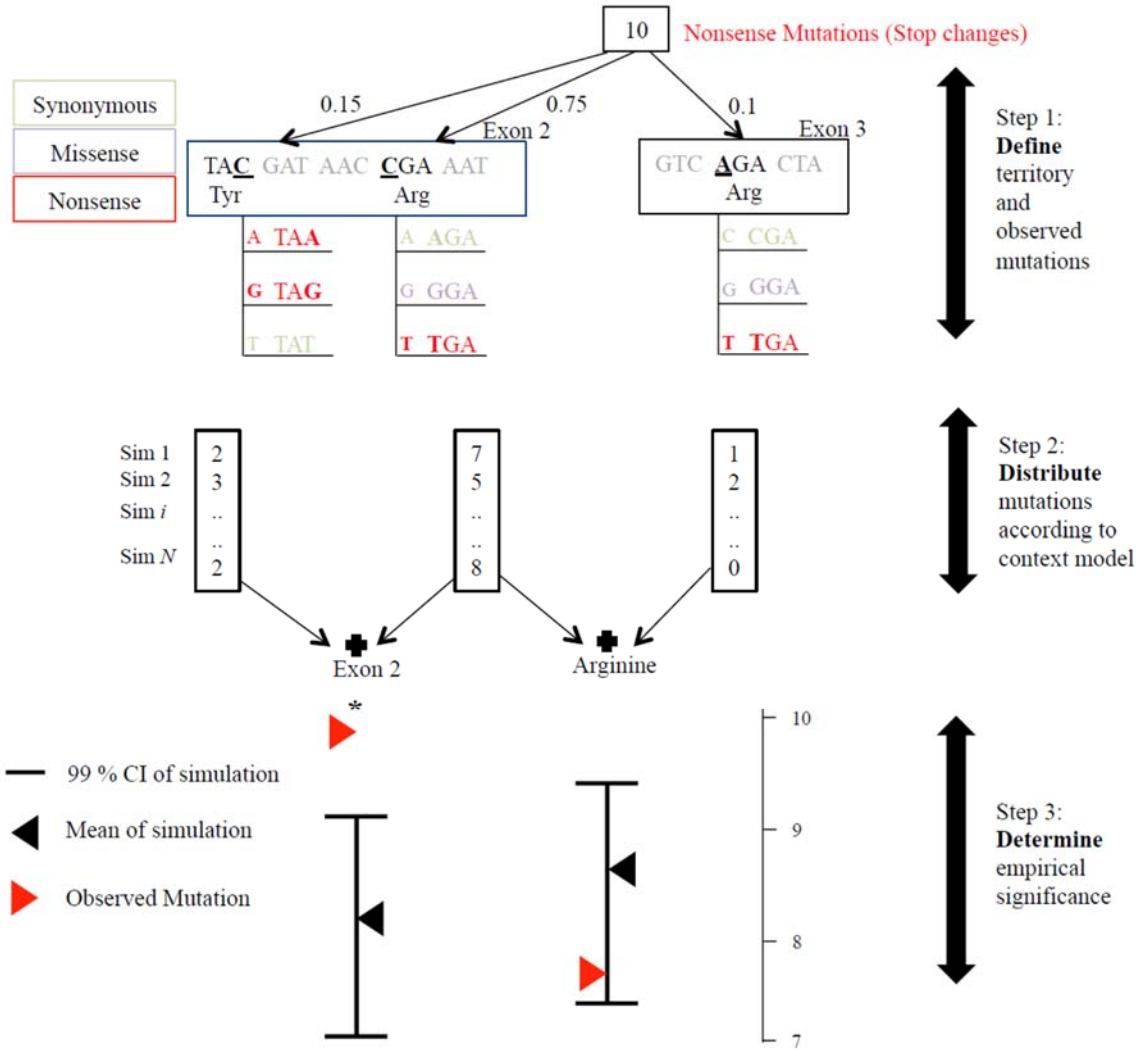


Figure 4.2 Detecting enrichment of nonsense mutations. Comparison of the overall observed number of mutations to the simulated frequency of nonsense mutations in both RB and ExAC datasets.

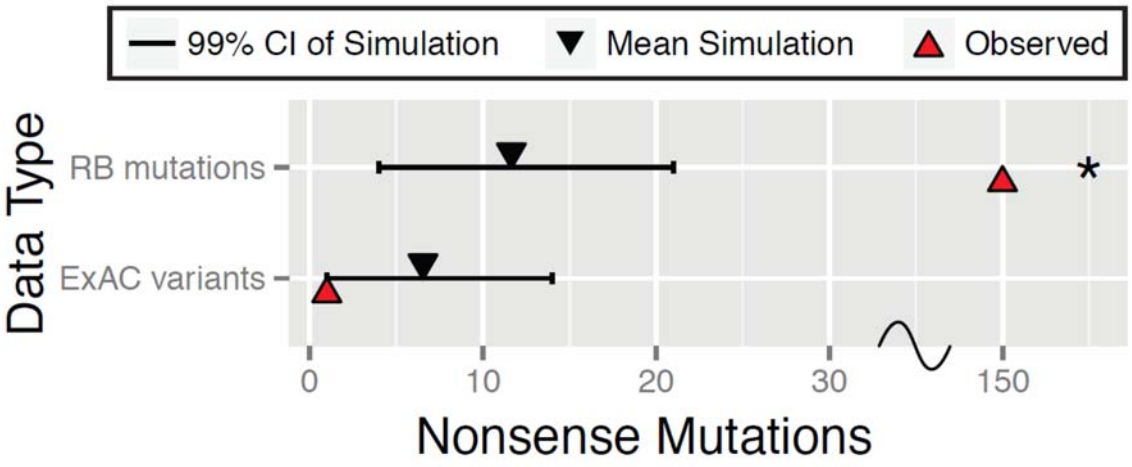


Figure 4.3 Detecting enrichment within nonsense mutations. Comparison of the observed number of mutations to the simulated frequency of nonsense mutations in RB, across exons 2 to 27. The asterisk (*) denotes that the observed number falls outside the 99% confidence interval (*i.e.*, $P < 0.01$). CI: Confidence Interval.

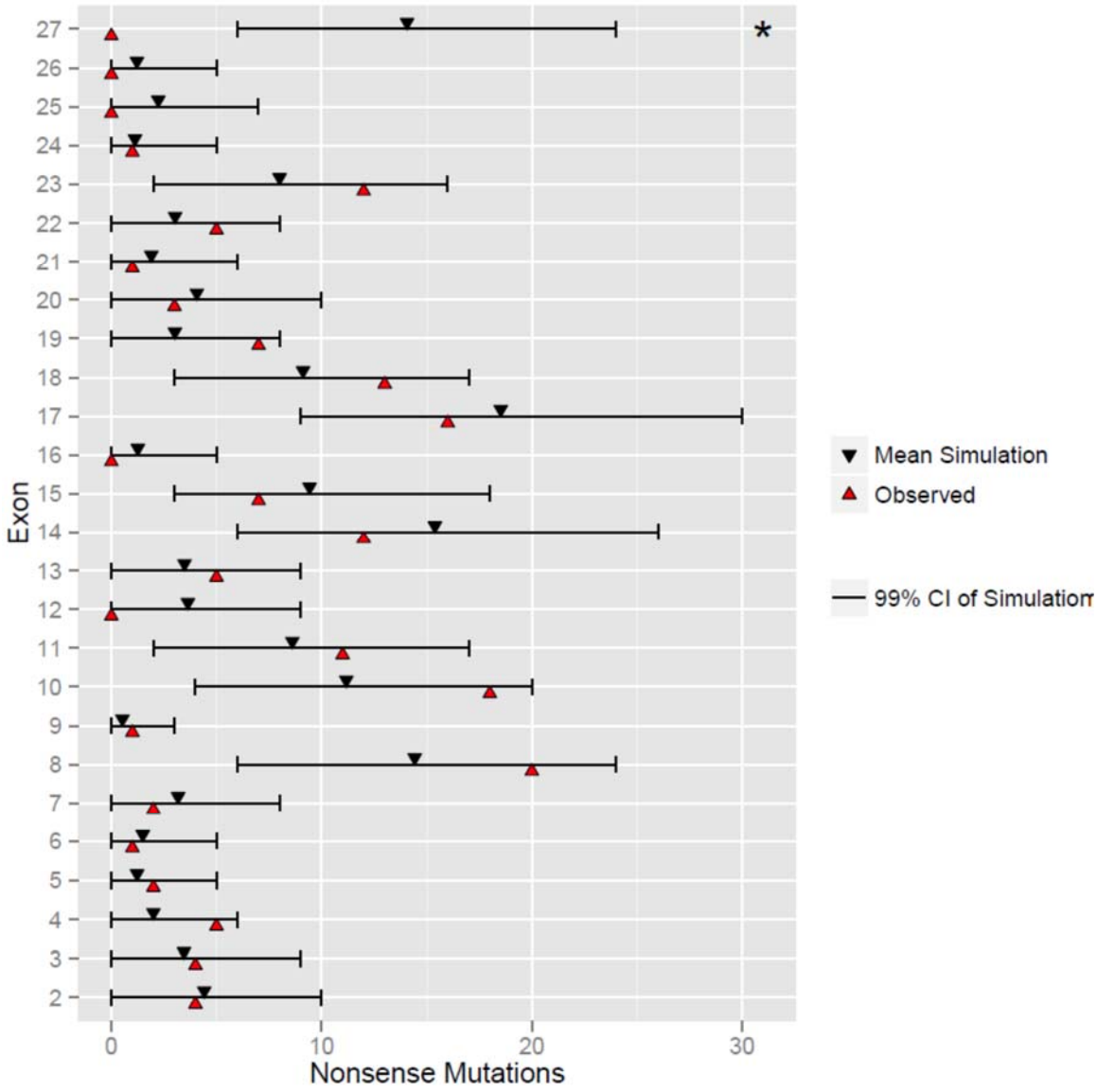


Figure 4.4 Detecting enrichment of splice mutations. Comparison of the overall observed number of mutations to the simulated frequency of essential splice and intronic mutations in both RB and ExAC datasets.

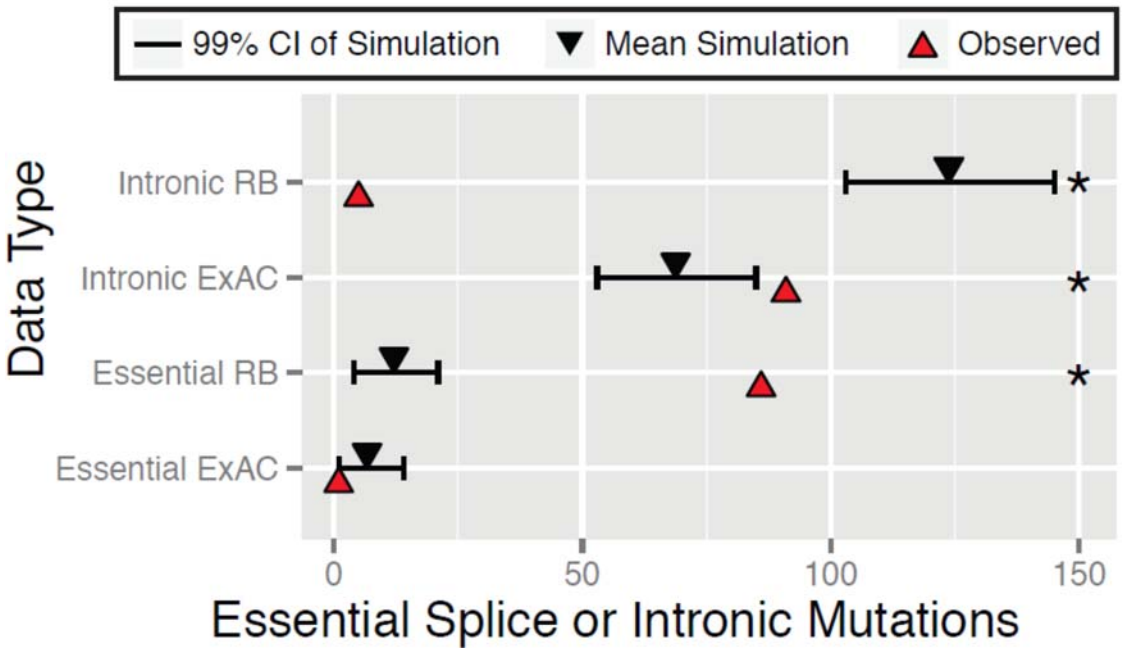


Figure 4.5 Detecting enrichment within essential splice acceptor mutations. Comparison of observed mutations and the simulated frequency of essential splice acceptor mutations in RB (99% CI) to find exon specific differential pathogenicity within essential splice mutations. Exons where the observed mutations are higher or lower than the 99% confidence interval of simulations are denoted by an asterisk (*).

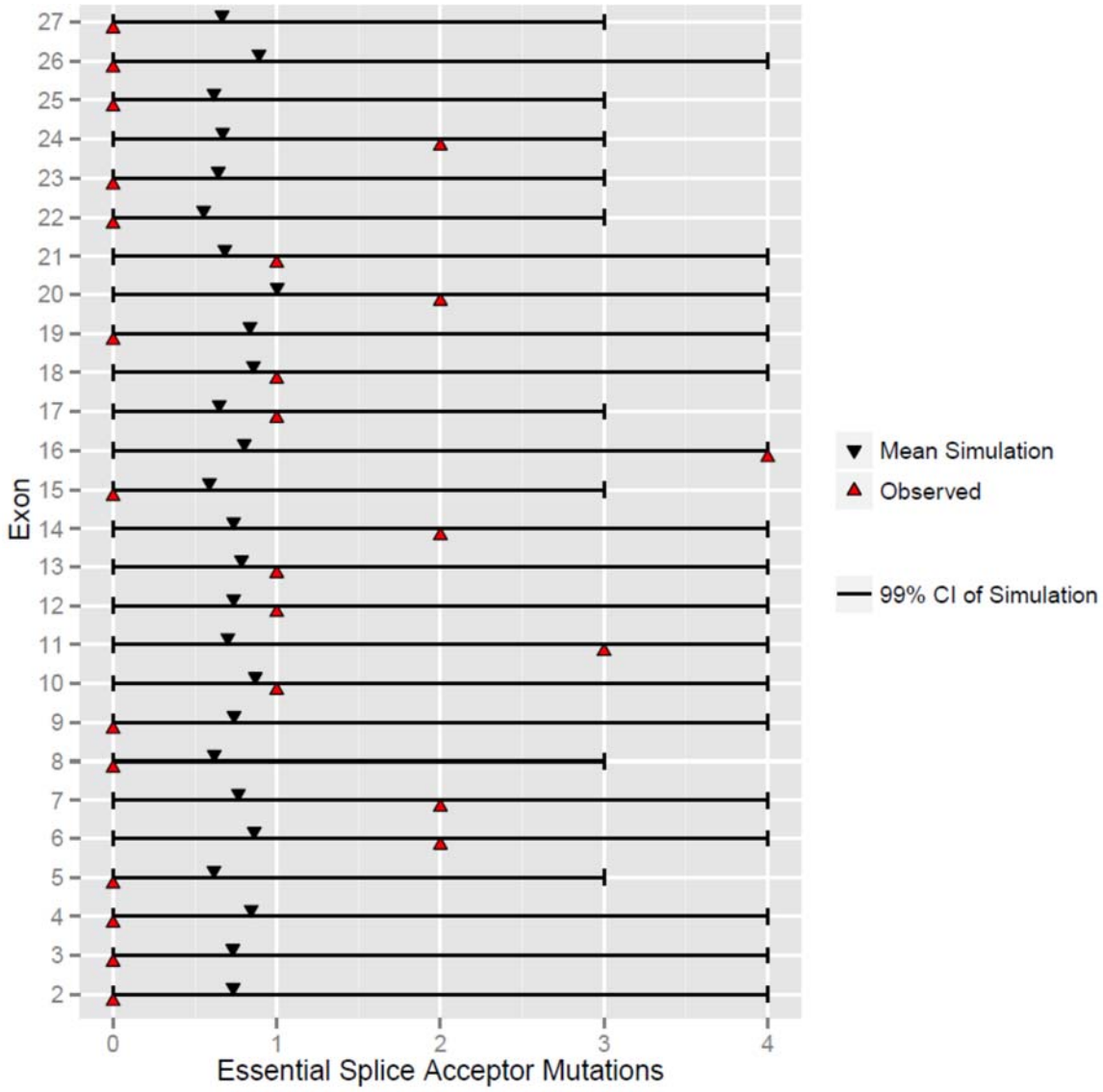


Figure 4.6 Detecting enrichment within essential splice donor mutations. Comparison of the observed number of mutations to the simulated frequency of essential splice donor mutations in RB, across exons 2 to 27. The asterisk (*) denotes that the observed number falls outside the 99% confidence interval (*i.e.*, P < 0.01). CI: Confidence Interval.

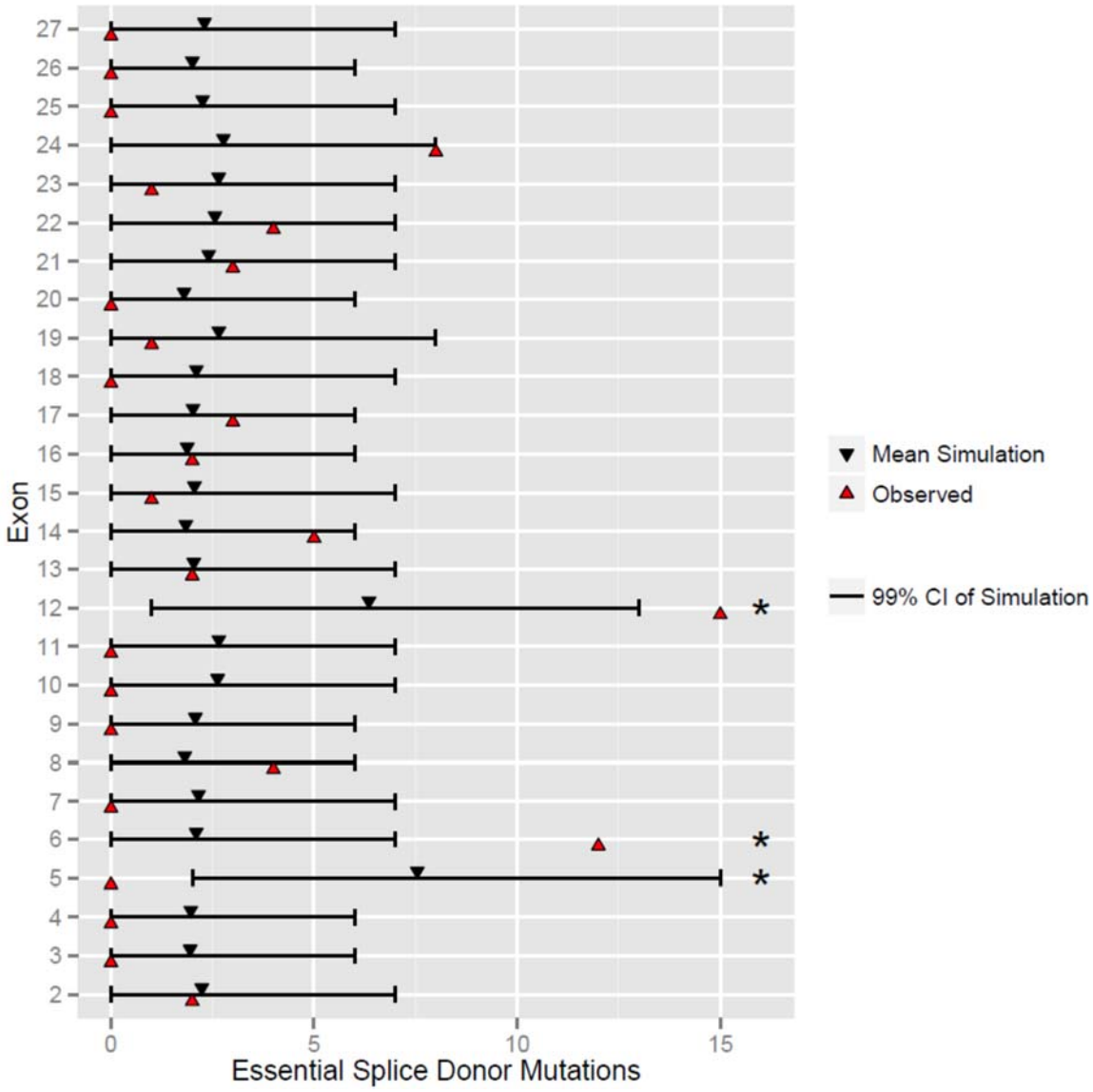


Figure 4.7 Splice mutations and their effect on codon structure. Donor splice mutations in Exons 5, 6 and 12, and their effect on codon structure. The codon structures are shown prior and after the donor splice mutation. The donor splice mutation results in exon skipping or deletion, but can also cause a frameshift mutation in certain cases.

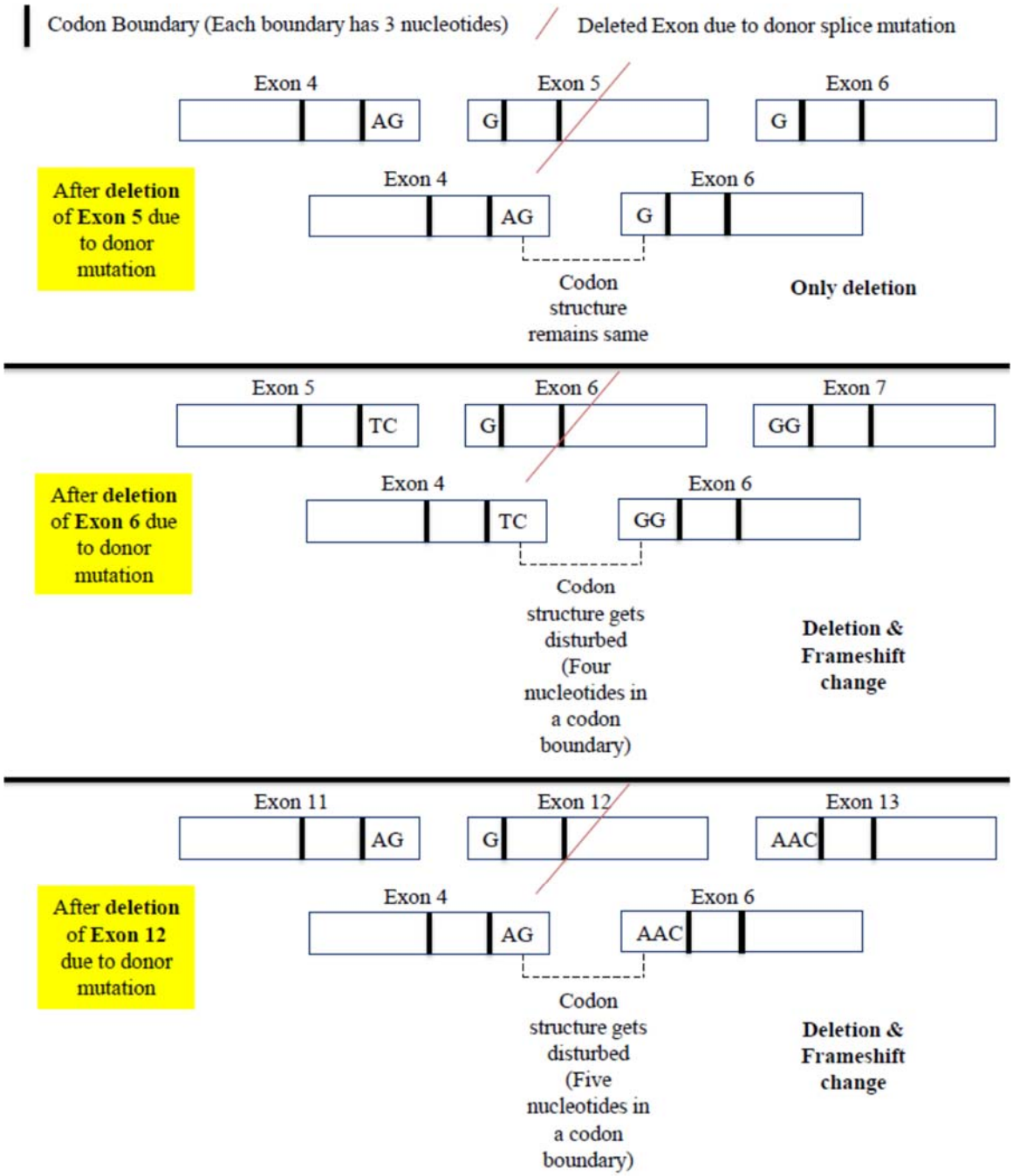


Figure 4.8 Detecting enrichment of missense mutations. Comparison of the overall observed number of mutations to the simulated frequency of missense mutations in both RB and ExAC datasets.

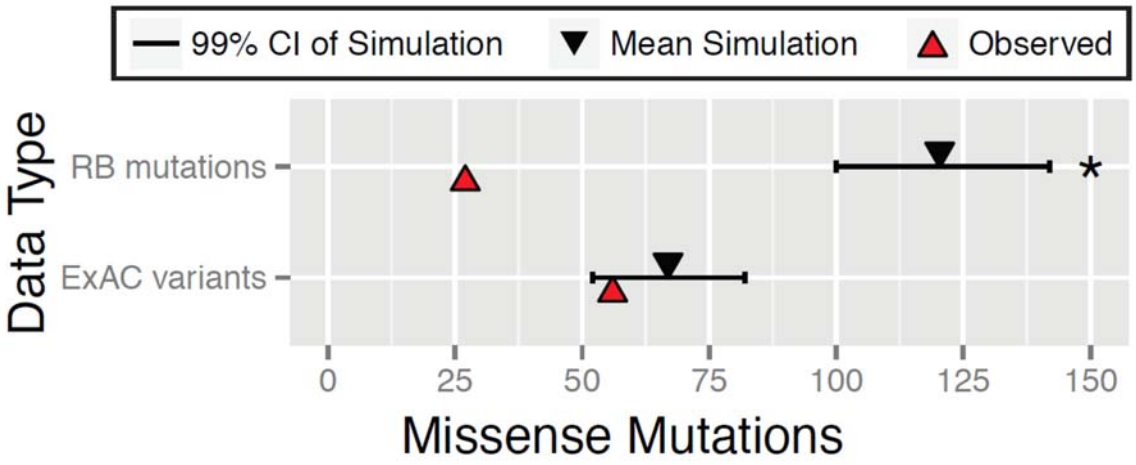


Figure 4.9 Detecting enrichment within missense mutations. Comparison of the observed number of mutations to the simulated frequency of missense mutations in RB, across exons 2 to 27.

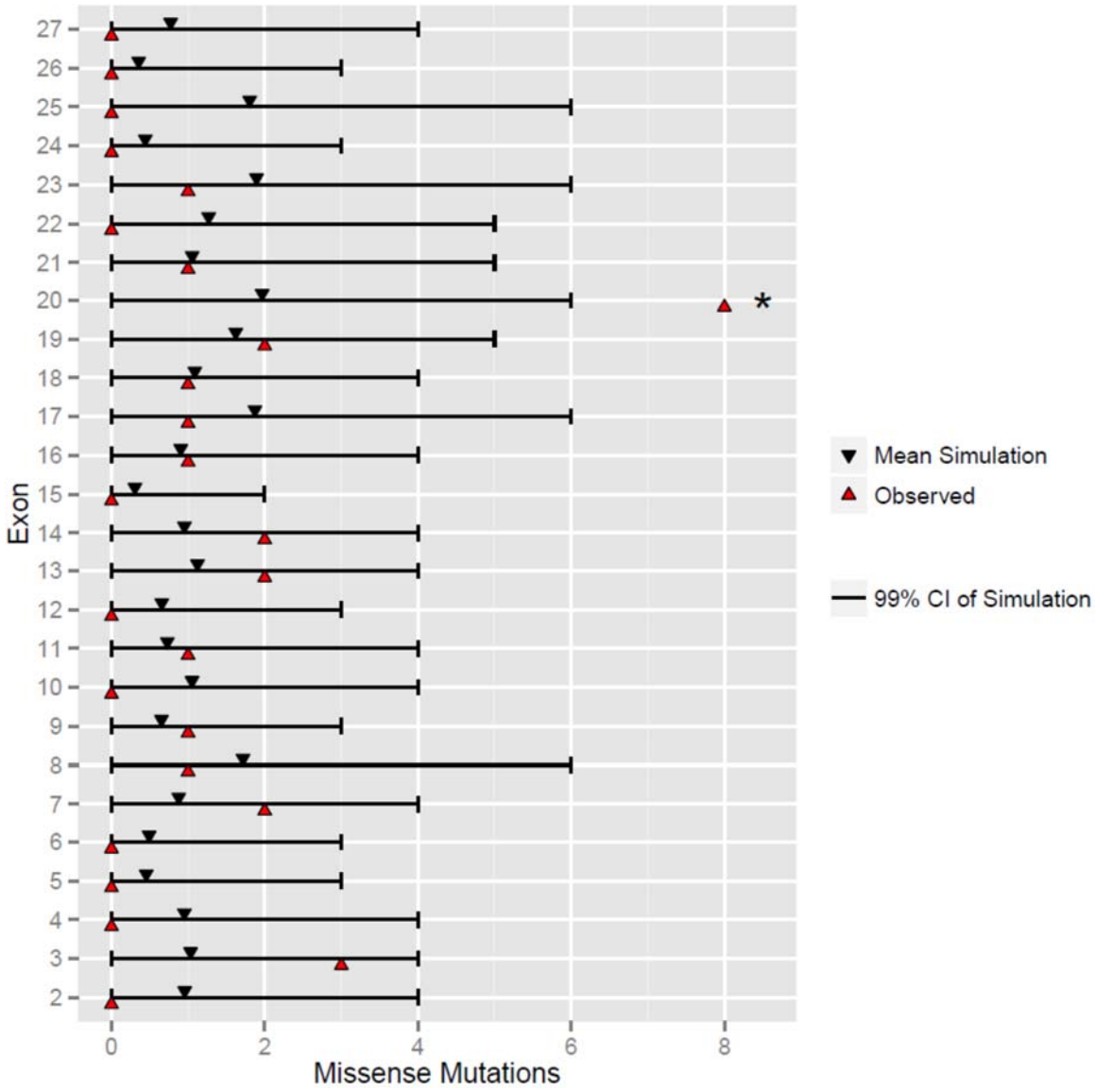


Figure 4.10 Detecting enrichment within missense mutations in the pocket domain. Comparison of the observed number of mutation to the simulated frequency of missense mutations over codons in the pocket domain of *RB1*. Here, a sliding window of 10 amino acids on either side of the codon was considered. Dotted line denotes the gap in the pocket domain.

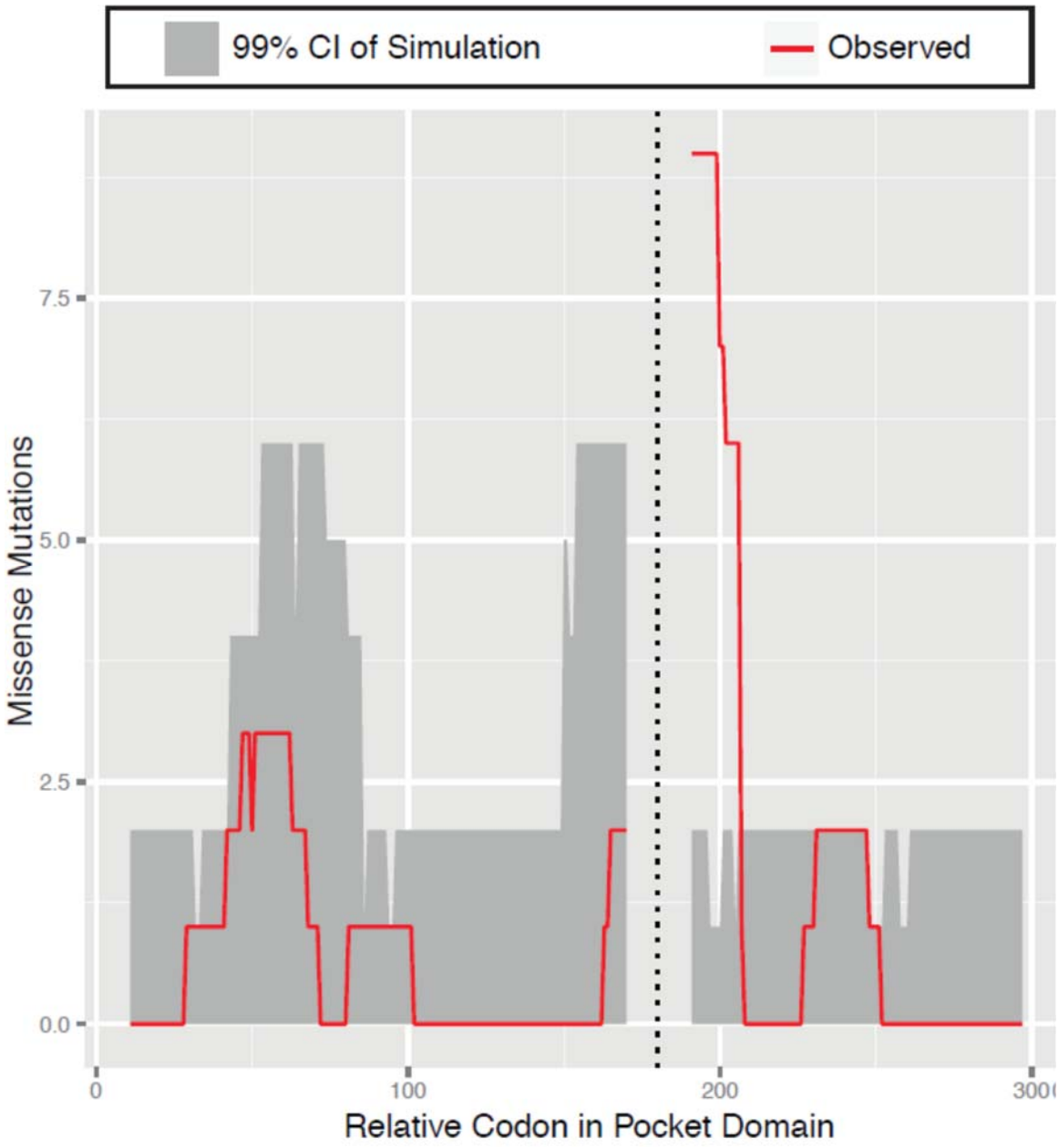


Figure 4.11 Relative rates of mutations across different categories. Comparison of the relative rates of different types of *de novo* mutations, normalized to the rate of nonsense mutations. Plotted is the mean of the ratio of observed number of mutations over expected based on the computational model. Mutational categories that have a different rate from the nonsense category ($P < 0.01$) are denoted by an asterisk (*). CI: Confidence Interval.

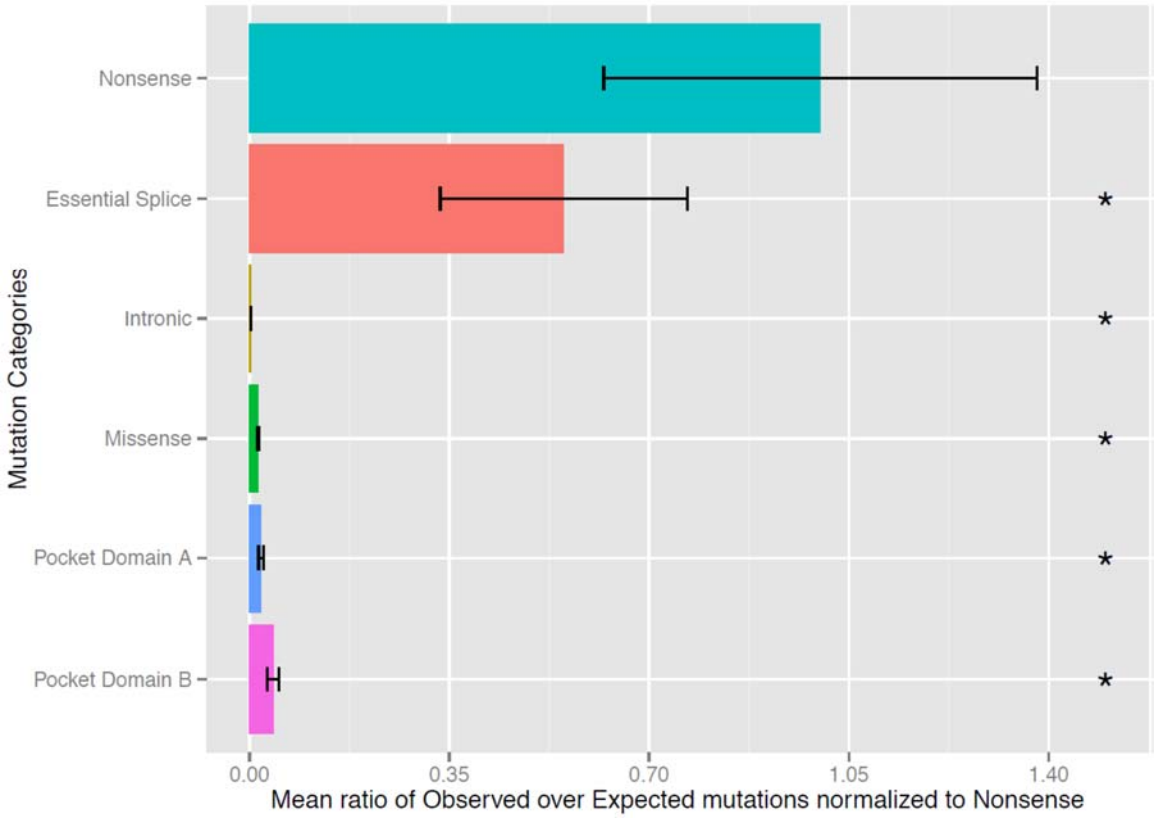


Figure 5.1 Correlation between de novo mutation rate estimates. Scatter plot of *de novo* mutation rate estimates at each heptanucleotide sequence context from polymorphism data from individuals of African and European ancestry.

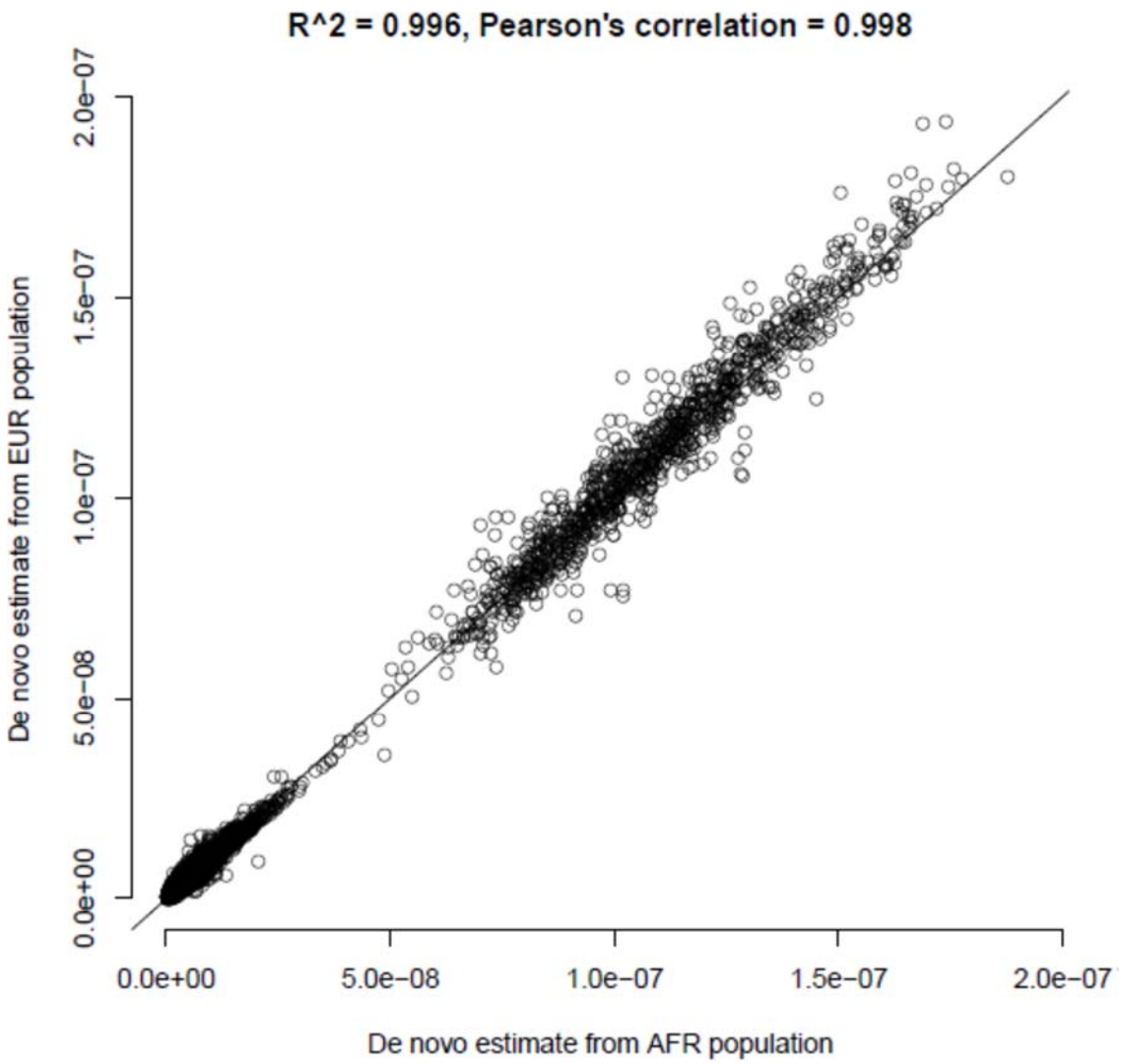


Figure 5.2 Predicted vs observed mutations at a CpG mutation motif. Boxplot of predicted *de novo* mutations, in a separate sample of germline mutations from whole genome sequencing of 78 trios at TACG motif changing to TATG. Predicted mutations are shown for heptanucleotide and trinucleotide sequence context based mutation rate estimates. Shown in dotted green line is the actual observed mutations at the original TACG motif changing to TATG.

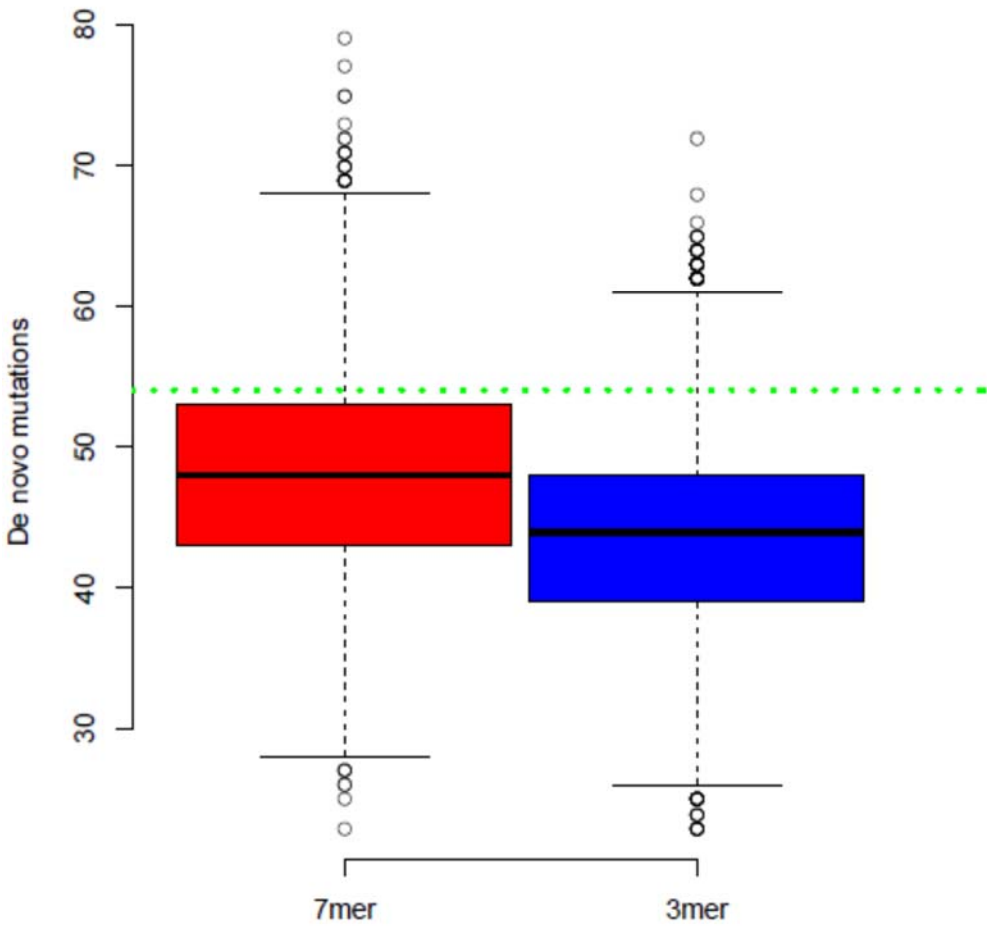


Figure 5.3 Predicted vs observed mutations at an ApT mutation motif. Boxplot of predicted *de novo* mutations, in a separate sample of germline mutations from whole genome sequencing of 78 trios at [C/T]CAAT motif changing to [C/T]CAGT. Predicted mutations are shown for heptanucleotide and trinucleotide sequence context based mutation rate estimates. Shown in dotted green line is the actual observed mutations at the original [C/T]CAAT motif changing to [C/T]CAGT.

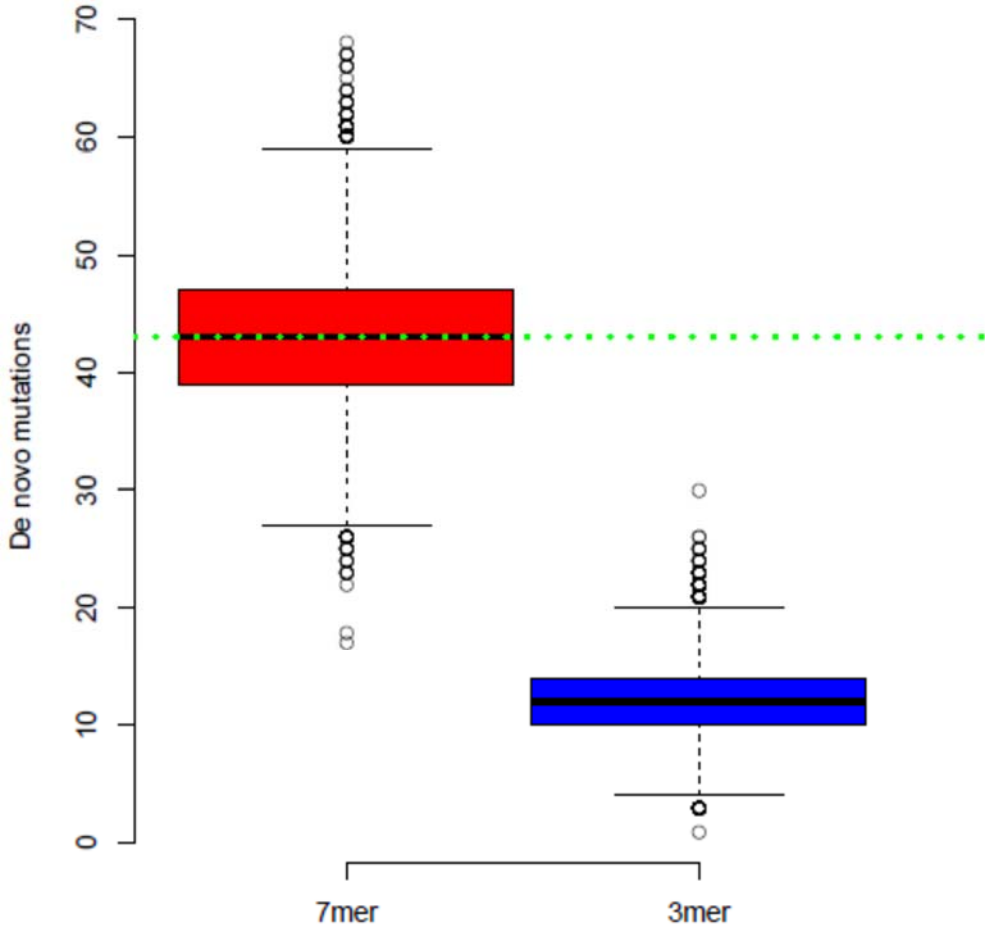


Figure 5.5 Variability in mutation rate between genes. Boxplot of predicted *de novo* mutation rate at each defined coding transcript using heptanucleotide sequence context based mutation rate estimates.

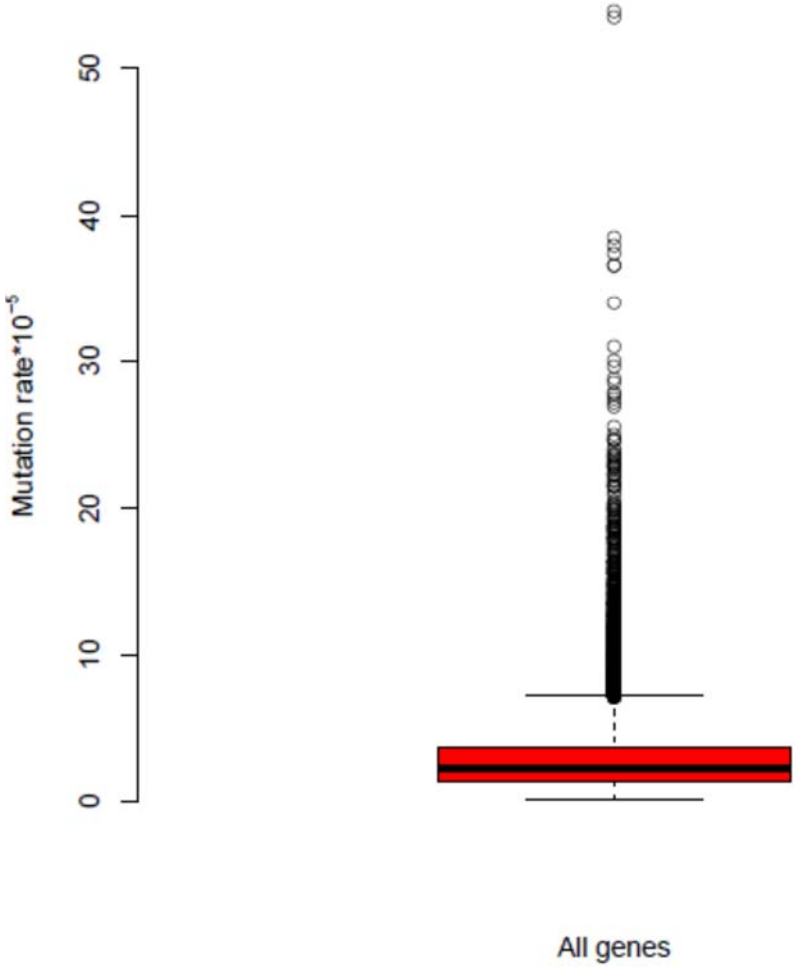


Figure 5.6 Variability in mutation rate between genes due to sequence context. Boxplot of predicted *de novo* mutation rate at each defined coding transcript using heptanucleotide sequence context based mutation rate estimates, further normalized by gene length. Shown in green is the average per base pair overall mutation rate in the coding region.

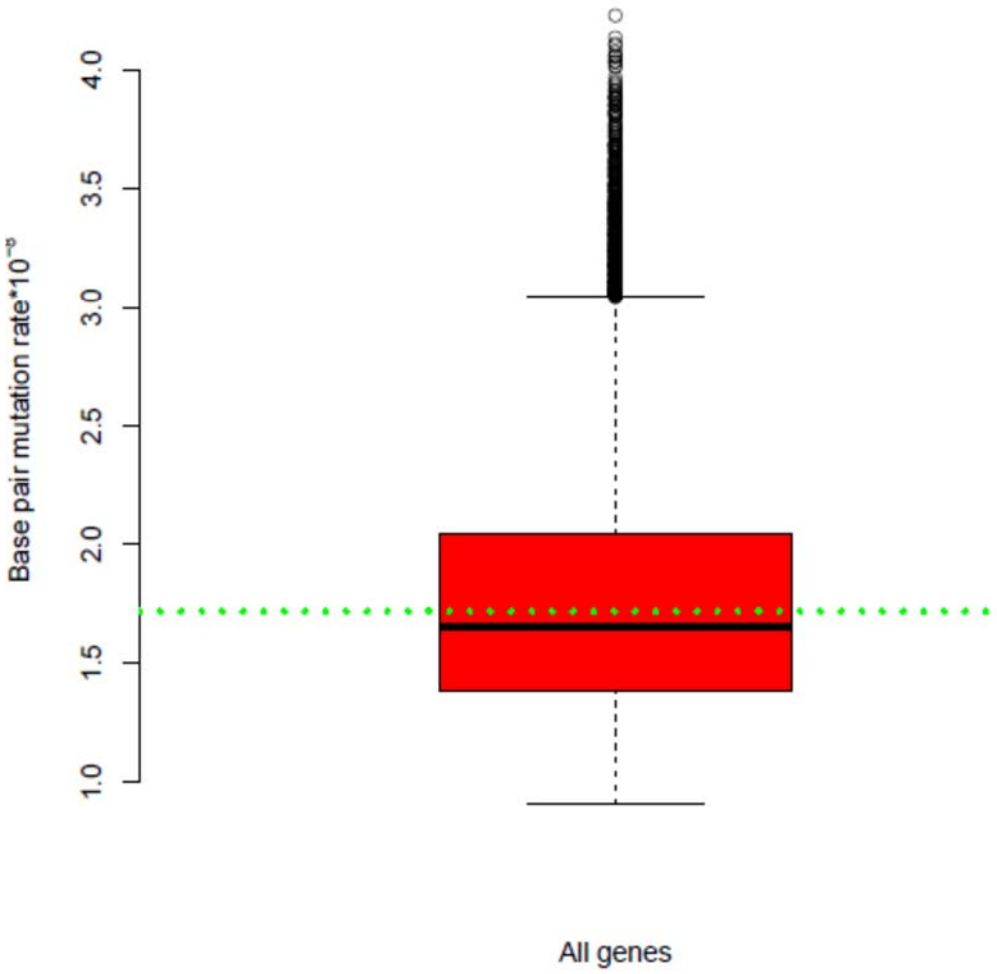
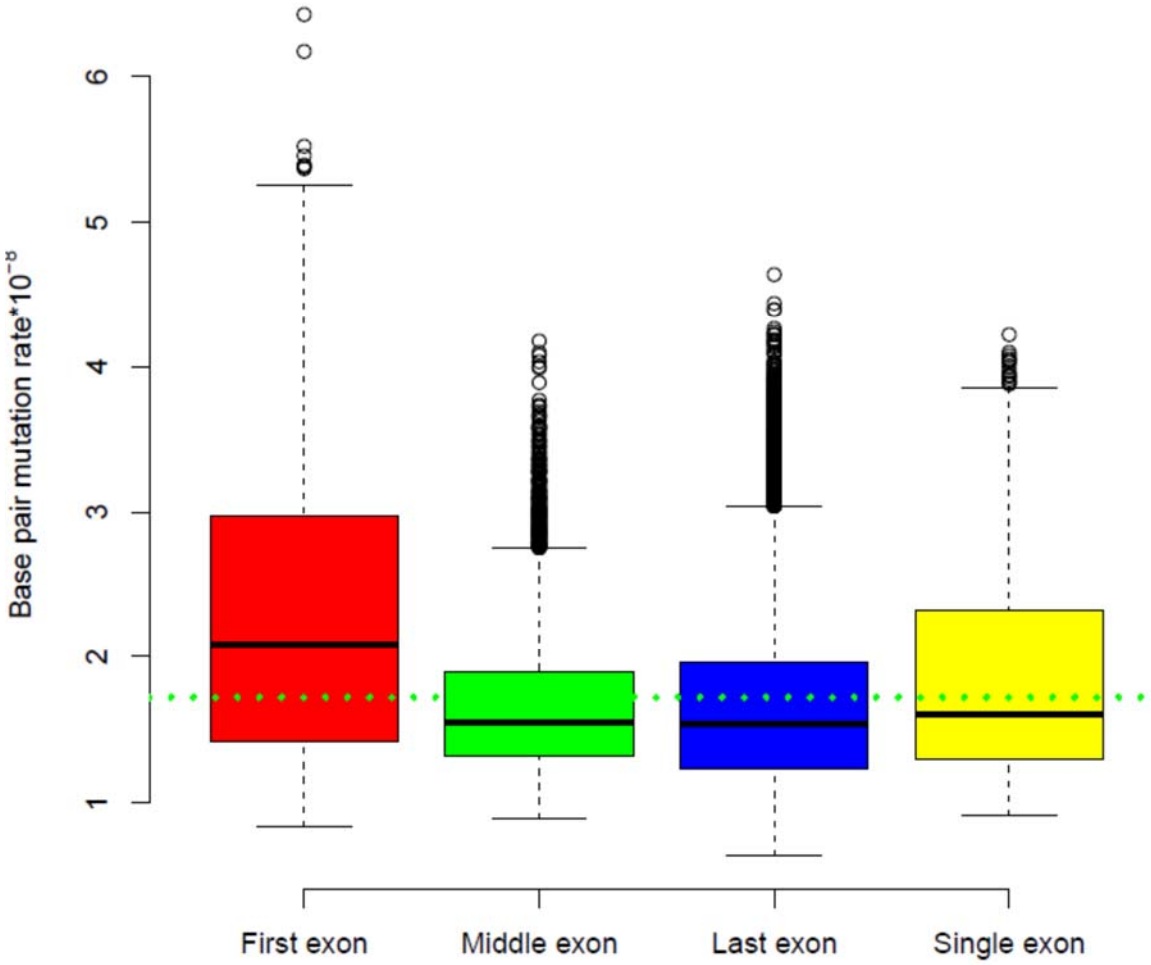


Figure 5.7 Variability in mutation rate across exons. Boxplot of predicted *de novo* mutation rate across the exon in each defined coding transcript using heptanucleotide sequence context based mutation rate estimates, further normalized by gene length. Shown in green is the average per base pair overall mutation rate in the coding region.


Base pair mutation rate*10^-8
1
2
3
4
5
6
First exon
Middle exon
Last exon
Single exon

SUPPLEMENTARY FILES

Supplementary File 2.1 Robustness of substitution probabilities. Pearson's correlation and Root Mean-Squared Error (RMSE) for substitution probabilities estimated from the training (all but the two listed chromosomes) and testing (the two listed chromosomes) sets from the intergenic non-coding genome. I present measurement for null (i.e., fixed rate), 1-mer, 3-mer, 5-mer and 7-mer models.

Supplementary File 2.2 LL comparison of different sequence context models. P-values for each likelihood ratio test comparing competing sequence context models (null, 1-mer, 3-mer, 5-mer and 7-mer), using all data from the intergenic non-coding genome. The matrix is symmetric, so "-" is presented where appropriate.

Supplementary File 2.3 Comparison of different sequence models on HapMap data. P-values and Bayes factor comparing sequence context models (1-mer, 3-mer, 5-mer and 7-mer) using all HapMap variant data from the intergenic non-coding genome.

Supplementary File 2.4 Bayes factor comparison of different sequence models. Natural logarithm of the approximate Bayes Factor comparing competing sequence context models (null, 1-mer, 3-mer, 5-mer and 7-mer), using all data from the intergenic non-coding genome. The matrix is symmetric, so "-" is presented where appropriate.

Supplementary File 2.5 Variability in substitution classes explained by different models. Stepwise regression model analysis for each substitution class on various models considered in the intergenic non-coding region. Data for the training phase was based on the collection even numbered chromosomes; data for the testing phase was based on odd numbered chromosomes. "# Features" denotes the features selected for that model. "AIC" is the Akaike Information Criterion, "MSE" represents Mean-squared Error; "adj-R2" is the adjusted R^2 from the model. The best performing model (lowest MSE after 8-fold cross validation) are highlighted in red.

Supplementary File 2.6 Sequence features and their effect on substitution probabilities. Aggregated sequence context features and their effect on the substitution probabilities for all classes of substitutions in the intergenic non-coding region. Order denotes the number of interacting nucleotides in the context. “BETA” indicates the regression coefficient for the sequence context for the given substitution class (i.e., A-to-C, A-to-G, etc.). “All_DIRXN” denotes the direction of effect for the feature on the substitution probability (+ indicates increase higher substitution probability, – indicates lower substitution probability). I present estimated values using (I) all data from 1KG, (II) data used in the training phase (all even-numbered chromosomes), and (III) data used for the testing phase (all odd-numbered chromosomes). (a) Substitution classes for sequence contexts outside of CpG sites (b) Substitution classes for sequences context including CpG sites (polymorphic 4th position is C and 5th position fixed at G).

Supplementary File 2.7 Nucleotide substitution probabilities in the noncoding region. Posterior probabilities of nucleotide substitution for all substitution classes within all 7-mer sequence contexts in the intergenic non-coding region for African, European and Asian populations groups (1KG project). The forward and reverse complementary sequences are presented for each probability.

Supplementary File 2.8 Comparison of different models across frequency spectrum. P-values and Bayes factor comparing sequence context models (1-mer, 3-mer, 5-mer and 7-mer) using all data from the intergenic non-coding genome. (a) Sequence context rates inferred from low frequency (1% and above MAF) variants from the 1000 genomes project (b) Sequence context rates inferred from rare (singletons and doubletons) variants from the 1000 genomes project.

Supplementary File 3.1 Nucleotide substitution probabilities in the coding region. Posterior probabilities of nucleotide substitution for all substitution classes within all 7-mer sequence contexts in coding region for African, European and Asian populations groups (1KG project). The forward and reverse complementary sequences are presented for each probability. The

corresponding amino acid changes associated with each substitution class within the 7-mer sequence context, as well as their reverse complements, are also listed in the table.

Supplementary File 3.2 Estimates of the variability in AA substitution probabilities. (a) Simulated and observed variance in nucleotide substitution probabilities grouped by type of amino acid replacement class. (b) Simulated and observed variance in nucleotide substitution probabilities, stratified for each possible types of amino acid replacement. Reported simulated values are based on 1,000,000 repetitions, based on a fixed rate model for each class of substitution.

Supplementary File 3.3 Gene Scores for functional intolerance. Gene scores and annotations for >16,000 transcripts in humans. I annotate each gene using Ensembl, attached to a specific transcript identifier. Columns 3 through 14 refer to the annotation attached to set membership (Essential, Ubiquitous, Immune, Olfactory, Keratin, Omim de novo, dominant, and haploinsufficient). Details and citations describing how each gene set was identified are presented in the Methods. The last three columns are gene scores calculated by my approach (for the African population), and various published methods.

Supplementary File 3.4 Amino Acid scores for specific AA functional intolerance. Amino acid tolerance scores for >16,000 transcripts in humans. These scores quantify the number of excess substitutions for each type of amino acid change relative to expected, with larger scores indicating fewer substitutions (intolerance) for that specific amino acid. Scores were developed using 1KG project data using the African group.

Supplementary File 4.1 De novo variants in RB1 gene from RB patients. All *de novo* germline variants in *RB1* gene of patients with RB. “gDNA position” is the nucleotide position in the GENBANK accession number L11910 of the gene.

Supplementary File 4.2 Singleton ExAC variants in RB1 gene. All ExAC variants in *RB1* gene that were considered in my analysis. “gDNA position” is the nucleotide position in the GENBANK accession number L11910 of the gene.

Supplementary File 4.3 *De novo* nonsense variants from an external dataset. All Nonsense variants in *RB1* gene from Onadim and Houdayer groups. “gDNA position” is the nucleotide position in the GENBANK accession number L11910 of the gene.

Supplementary File 5.1 heptanucleotide mutation rate estimates from AFR polymorphism data. *De novo* mutation rate estimate at each heptanucleotide sequence context (“alpha”) to another (“beta”) from polymorphism data of individuals from African ancestry.

Supplementary File 5.2 heptanucleotide mutation rate estimates from EUR polymorphism data. *De novo* mutation rate estimate at each heptanucleotide sequence context (“alpha”) to another (“beta”) from polymorphism data of individuals from European ancestry.

Supplementary File 5.3 trinucleotide de novo mutation rate estimates. *De novo* mutation rate estimates at each trinucleotide sequence context (“alpha”) to another (“beta”) from polymorphism data of individuals from African ancestry.

Supplementary File 5.4 1mer+CpG de novo mutation rate estimates. *De novo* mutation rate estimate at each 2mer sequence context (“alpha”) to another (“beta”) from polymorphism data of individuals from African ancestry. The first position at each 2mer context is the base which is mutated. The rates reported here use the 1mer *de novo* mutation rate estimates for all changes, except for C-to-T mutation at CpG site.

BIBLIOGRAPHY

1. Rosenberg SM: **Evolving responsively: adaptive mutation**. *Nat Rev Genet* 2001, **2**:504–515.

2. Griffiths AJ, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM: **Sources of variation**. 2000.

3. Hodgkinson A, Eyre-Walker A: **Variation in the mutation rate across mammalian genomes.** *Nat Rev Genet* 2011, **12**:756–66.

4. Baer CF, Miyamoto MM, Denver DR: **Mutation rate variation in multicellular eukaryotes: causes and consequences**. *Nat Rev Genet* 2007, **8**:619–631.

5. Marusyk A, Almendro V, Polyak K: **Intra-tumour heterogeneity: a looking glass for cancer?** *Nat Rev Cancer* 2012, **12**:323–334.

6. Veltman JA, Brunner HG: **De novo mutations in human genetic disease.** *Nat Rev Genet* 2012, **13**:565–75.

7. Martincorena I, Campbell PJ, Stratton MR, Campbell PJ, Futreal PA, Loeb LA, Harris CC, Knudson AG, Cairns J, Nowell PC, Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ, Greenman CD, Jia M, Latimer C, Teague JW, Lau KW, Burton J, Quail MA, Swerdlow H, Churcher C, Natrajan R, et al.: **Somatic mutation in cancer and normal cells.** *Science* 2015, **349**:1483–9.

8. Ségurel L, Wyman MJ, Przeworski M: **Determinants of mutation rate variation in the human germline.** *Annu Rev Genomics Hum Genet* 2014, **15**:47–70.

9. Scally A, Durbin R: **Revising the human mutation rate: implications for understanding human evolution.** *Nat Rev Genet* 2012, **13**:745–53.

10. Sequencing T: **Initial sequence of the chimpanzee genome and comparison with the human genome**. *Nature* 2005, **437**:69–87.

11. Kimura M: *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press; 1983.

12. Jukes TH, Kimura M: **Evolutionary constraints and the neutral theory**. *J Mol Evol* 1984, **21**:90–92.

13. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491**:56–65.

14. Schaibley VM, Zawistowski M, Wegmann D, Ehm MG, Nelson MR, St Jean PL, Abecasis GR, Novembre J, Zöllner S, Li JZ: **The influence of genomic context on mutation patterns in the human genome inferred from rare variants.** *Genome Res* 2013, **23**:1974–84.

15. Ewens WJ: *Mathematical Population Genetics*. Springer; 2004.

16. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WSW, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K: **Rate of de novo mutations and the importance of father's age to disease risk.** *Nature* 2012, **488**:471–5.

17. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, van Duijn CM, Swertz M, Wijmenga C, van Ommen G, Slagboom PE, Boomsma DI, Ye K, Guryev V, Arndt PF, Kloosterman WP, de Bakker PIW, Sunyaev SR, Sunyaev SR: **Genome-wide patterns and properties of de novo mutations in humans**. *Nat Genet* 2015, **47**:822–826.

18. Goldmann JM, Wong WSW, Pinelli M, Farrah T, Bodian D, Stittrich AB, Glusman G, Vissers LELM, Hoischen A, Roach JC, Vockley JG, Veltman JA, Solomon BD, Gilissen C, Niederhuber JE: **Parent-of-origin-specific signatures of de novo mutations**. *Nat Genet* 2016, **48**:935–939.

19. Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA: **Differential relationship of DNA replication timing to different forms of human mutation and variation.** *Am J Hum Genet* 2012, **91**:1033–40.

20. Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov G V, Mirkin SM, Sunyaev SR: **Human mutation rate associated with DNA replication timing**. *Nat Genet* 2009, **41**:393–395.

21. Ehrlich M, Wang RY: **5-Methylcytosine in eukaryotic DNA.** *Science* 1981, **212**:1350–7.

22. Ohno M, Sakumi K, Fukumura R, Furuichi M, Iwasaki Y, Hokama M, Ikemura T, Tsuzuki T, Gondo Y, Nakabeppu Y, Kong A, Keightley PD, Xue Y, Casals F, Bertranpetit J, Ohno M, Shibutani S, Takeshita M, Grollman AP, Maki H, Sekiguchi M, Michaels ML, Cruz C, Grollman AP, Miller JH, Sakumi K, Radicella JP, Dherin C, Desmaze C, Fox MS, et al.: **8-oxoguanine causes spontaneous de novo germline mutations in mice**. *Sci Rep* 2014, **4**:471–475.

23. Helleday T, Eshtad S, Nik-Zainal S: **Mechanisms underlying mutational signatures in human cancers**. *Nat Rev Genet* 2014, **15**:585–598.

24. Ciccia A, Elledge SJ, Adamo A, Collis SJ, Adelman CA, Silva N, Horejsi Z, Ward JD, Martinez-Perez E, Boulton SJ, Volpe A La, Ahel I, Ahel D, Matsusaka T, Clark AJ, Pines J, Boulton SJ, West SC, Ahel D, Horejsí Z, Wiechens N, Polo SE, Garcia-Wilson E, Ahel I, Flynn H, Skehel M, West SC, Jackson SP, al. et, Andersen SL, et al.: **The DNA Damage Response: Making It Safe to Play with Knives**. *Mol Cell* 2010, **40**:179–204.

25. Crow JF: **The origins, patterns and implications of human spontaneous mutation.** *Nat Rev Genet* 2000, **1**:40–7.

26. Roberts ' JD, Kunke12 TA: **Fidelity of DNA Replication Discrimination Steps in a Polymerization Cycle**. .

27. Yakovchuk P, Protozanova E, Frank-Kamenetskii MD: **Base-stacking and base-pairing contributions into thermal stability of the DNA double helix.** *Nucleic Acids Res* 2006, **34**:564–74.

28. Nachman MW, Crowell SL: **Estimate of the Mutation Rate per Nucleotide in Humans**. *Genetics* 2000, **156**:297–304.

29. Aggarwala V, Voight BF: **An expanded sequence context model broadly explains variability in polymorphism levels across the human genome**. *Nat Genet* 2016, **48**:349–355.

30. Hwang DG, Green P: **Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution.** *Proc Natl Acad Sci U S A* 2004, **101**:13994–4001.

31. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnström K, Mallick S, Kirby A, Wall DP, MacArthur DG, Gabriel SB, DePristo M, Purcell SM, Palotie A, Boerwinkle E, Buxbaum JD, Cook EH, Gibbs RA, Schellenberg GD, Sutcliffe JS, Devlin B, Roeder K, Neale BM, Daly MJ: **A framework for the interpretation of de novo mutation in human disease**. *Nat Genet* 2014, **46**:944–950.

32. Lawrence MS, Stojanov P, Polak P, Kryukov G V, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortés ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, et al.: **Mutational heterogeneity in cancer and the search for new cancer-associated genes.** *Nature* 2013, **499**:214–8.

33. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, Wu W, Corominas R, Peoples A, Koren A, Gore A, Kang S, Lin GN, Estabillo J, Gadomski T, Singh B, Zhang K, Akshoomoff N, Corsello C, McCarroll S, Iakoucheva LM, Li Y, Wang J, Sebat J: **Whole-genome sequencing in autism identifies hot spots for de novo germline mutation.** *Cell* 2012, **151**:1431–42.

34. Pfeifer GP, You Y-H, Besaratinia A: **Mutations induced by ultraviolet light**. *Mutat Res Mol Mech Mutagen* 2005, **571**:19–31.

35. Ruggeri B, Dirado M, Zhang SY, Bauer B, Goodrowt T, Klein-Szanto AJP: **Benzo[a]pyrene-induced murine skin tumors exhibit frequent and characteristic G to T mutations in the p53 gene**. *Med Sci* 1993, **90**:1013–1017.

36. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinsk M, Jäger N, Jones DTW, Jones D, Knappskog S, Kool M, et al.: **Signatures of mutational processes in human cancer**. *Nature* 2013, **500**:415–421.

37. Fousteri M, Mullenders LHF: **Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects.** *Cell Res* 2008, **18**:73–84.

38. McVicker G, Green P: **Genomic signatures of germline gene expression.** *Genome Res* 2010, **20**:1503–11.

39. Li WH, Wu CI, Luo CC: **Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications.** *J Mol Evol* 1984, **21**:58–71.

40. Charlesworth B, Morgan MT, Charlesworth D: **The effect of deleterious mutations on neutral molecular variation.** *Genetics* 1993, **134**:1289–303.

41. Charlesworth B, Aguadé M, Miyashita N, Langley CH, Andolfatto P, Andolfatto P, Przeworski M, Arguello JR, Zhang Y, Kado T, Fan CZ, Zhao RP, Ashburner M, Golic KG, Hawley RS, Bachtrog D, Hom E, Wong KM, Maside X, Jong P De, Barraclough TG, Fontaneto D, Ricci C, Herniou EA, Bartolomé C, Charlesworth B, Barton NH, Barton NH, Barton NH, Etheridge AM, et al.: **The effects of deleterious mutations on evolution at linked sites.** *Genetics* 2012, **190**:5–22.

42. Eyre-Walker A, Keightley PD: **The distribution of fitness effects of new mutations.** *Nat Rev Genet* 2007, **8**:610–8.

43. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko J V, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES: **Detecting recent positive selection in the human genome from haplotype structure.** *Nature* 2002, **419**:832–7.

44. Charlesworth D: **Balancing Selection and Its Effects on Sequences in Nearby Genome Regions**. *PLoS Genet* 2006, **2**:e64.

45. Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M: **Widespread purifying selection at polymorphic sites in human protein-coding loci**. *Proc Natl Acad Sci* 2003, **100**:15754–15757.

46. Ward LD, Kellis M: **Evidence of abundant purifying selection in humans for recently acquired regulatory functions.** *Science* 2012, **337**:1675–8.

47. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M: **Natural selection on genes that underlie human disease susceptibility.** *Curr Biol* 2008, **18**:883–9.

48. Cai JJ, Borenstein E, Chen R, Petrov DA: **Similarly Strong Purifying Selection Acts on**

Human Disease Genes of All Evolutionary Ages. *Genome Biol Evol* 2010, **1**:131–144.

49. Maher MC, Uricchio LH, Torgerson DG, Hernandez RD: **Population Genetics of Rare Variants and Complex Diseases**. *Hum Hered* 2012, **74**:118–128.

50. Vasseur E, Quintana-Murci L: **The impact of natural selection on health and disease: uses of the population genetics approach in humans**. *Evol Appl* 2013, **6**:596–607.

51. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, Barrett JC, Biesecker LG, Conrad DF, Cooper GM, Cox NJ, Daly MJ, Gerstein MB, Goldstein DB, Hirschhorn JN, Leal SM, Pennacchio LA, Stamatoyannopoulos JA, Sunyaev SR, Valle D, Voight BF, Winckler W, Gunter C: **Guidelines for investigating causality of sequence variants in human disease**. *Nature* 2014, **508**:469–476.

52. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Ercument Cicek A, Kou Y, Liu L, Fromer M, Walker S, Singh T, Klei L, Kosmicki J, Fu S-C, Aleksic B, Biscaldi M, Bolton PF, Brownfeld JM, Cai J, Campbell NG, Carracedo A, Chahrour MH, Chiocchetti AG, Coon H, Crawford EL, Crooks L, Curran SR, Dawson G, Duketis E, Fernandez BA, et al.: **Synaptic, transcriptional and chromatin genes disrupted in autism**. *Nature* 2014, **515**:209–215.

53. Homsy J, Zaidi S, Shen Y, Ware JS, Samocha KE, Karczewski KJ, DePalma SR, McKean D, Wakimoto H, Gorham J, Jin SC, Deanfield J, Giardini A, Porter GA, Kim R, Bilguvar K, López-Giráldez F, Tikhonova I, Mane S, Romano-Adesman A, Qi H, Vardarajan B, Ma L, Daly M, Roberts AE, Russell MW, Mital S, Newburger JW, Gaynor JW, Breitbart RE, et al.: **De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies.** *Science* 2015, **350**:1262–6.

54. Yang Z, Bielawski JP: **Statistical methods for detecting molecular adaptation**. *Trends Ecol Evol* 2000, **15**:496–503.

55. Hurst LD: **The Ka/Ks ratio: diagnosing the form of sequence evolution**. *Trends Genet* 2002, **18**:486–487.

56. Charlesworth B: **Stabilizing selection, purifying selection, and mutational bias in finite populations.** *Genetics* 2013, **194**:955–71.

57. Sunyaev S, Kondrashov FA, Bork P, Ramensky V: **Impact of selection, mutation rate and genetic drift on human genetic variation.** *Hum Mol Genet* 2003, **12**:3325–30.

58. Stenson PD, Mort M, Ball E V, Shaw K, Phillips A, Cooper DN: **The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.** *Hum Genet* 2014, **133**:1–9.

59. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J: **Mutations: Types and Causes**. 2000.

60. Vogel F: **Genetics of retinoblastoma.** *Hum Genet* 1979, **52**:1–54.

61. Parkin DM, Stiller CA, Draper GJ, Bieber CA: **The international incidence of childhood cancer.** *Int J cancer* 1988, **42**:511–20.

62. Dimaras H, Corson TW, Cobrinik D, White A, Zhao J, Munier FL, Abramson DH, Shields CL, Chantada GL, Njuguna F, Gallie BL, Knudson AG, Hanahan D, Weinberg RA, Dimaras H, Dimaras H, Seregard S, Lundell G, Svedberg H, Kivela T, Kivela T, Broaddus E, Topham A, Singh AD, Nyamori JM, Kimani K, Njuguna MW, Dimaras H, MacCarthy A, Krishna SM, et al.: **Retinoblastoma**. *Nat Rev Dis Prim* 2015, **1**:15021.

63. Lohmann DR, Gallie BL: *Retinoblastoma*. University of Washington, Seattle; 1993.

64. Knudson AG: **Mutation and cancer: statistical study of retinoblastoma.** *Proc Natl Acad*

Sci U S A 1971, **68**:820–3.

65. Abramson DH, Beaverson K, Sangani P, Vora RA, Lee TC, Hochberg HM, Kirszrot J, Ranjithan M: **Screening for retinoblastoma: presenting signs as prognosticators of patient and ocular survival.** *Pediatrics* 2003, **112**(6 Pt 1):1248–55.

66. Nichols KE, Houseknecht MD, Godmilow L, Bunin G, Shields C, Meadows A, Ganguly A: **Sensitive multistep clinical molecular screening of 180 unrelated individuals with retinoblastoma detects 36 novel mutations in the RB1 gene.** *Hum Mutat* 2005, **25**:566–74.

67. Lohmann DR, Brandt B, Höpping W, Passarge E, Horsthemke B: **The spectrum of RB1 germ-line mutations in hereditary retinoblastoma.** *Am J Hum Genet* 1996, **58**:940–9.

68. Valverde JR, Alonso J, Palacios I, Pestaña Á: **RB1 gene mutation up-date, a meta-analysis based on 932 reported mutations available in a searchable database**. *BMC Genet* 2005, **6**.

69. Lohmann DR, Gerick M, Brandt B, Oelschläger U, Lorenz B, Passarge E, Horsthemke B: **Constitutional RB1-gene mutations in patients with isolated unilateral retinoblastoma.** *Am J Hum Genet* 1997, **61**:282–94.

70. Smith T, Ho G, Christodoulou J, Price EA, Onadim Z, Gauthier-Villars M, Dehainault C, Houdayer C, Parfait B, van Minkelen R, Lohman D, Eyre-Walker A: **Extensive Variation in the Mutation Rate Between and Within Human Genes Associated with Mendelian Disease.** *Hum Mutat* 2016, **37**:488–94.

71. Onadim Z, Hogg A, Baird PN, Cowell JK: **Oncogenic point mutations in exon 20 of the RB1 gene in families showing incomplete penetrance and mild expression of the retinoblastoma phenotype.** *Proc Natl Acad Sci U S A* 1992, **89**:6177–81.

72. Otterson GA, Chen W d, Coxon AB, Khleif SN, Kaye FJ: **Incomplete penetrance of familial retinoblastoma linked to germ-line mutations that result in partial loss of RB function.** *Proc Natl Acad Sci U S A* 1997, **94**:12036–40.

73. Lek M, Karczewski KJ, Minikel E V, samocha KE, Banks E, Fennell timothy, O anne H, Ware J, Hill andrew J, cummings BB, tukiainen taru, Birnbaum DP, Kosmicki J, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, gauthier L, goldstein J, gupta N, Howrigan D, Kiezun adam, Kurki M, Rivas M, Ruano-Rubio V, Rose samuel, Ruderfer DM, shakir K, et al.: **Analysis of protein-coding genetic variation in 60,706 humans**. *Nat Publ Gr* 2016, **536**.

74. Huguet G, Ey E, Bourgeron T: **The Genetic Landscapes of Autism Spectrum Disorders**. *Annu Rev Genomics Hum Genet* 2013, **14**:191–213.

75. Elsabbagh M, Divan G, Koh Y-J, Kim YS, Kauchali S, Marcín C, Montiel-Nava C, Patel V, Paula CS, Wang C, Yasamy MT, Fombonne E: **Global prevalence of autism and other pervasive developmental disorders.** *Autism Res* 2012, **5**:160–79.

76. Werling DM, Geschwind DH: **Sex differences in autism spectrum disorders.** *Curr Opin Neurol* 2013, **26**:146–53.

77. Jeste SS, Geschwind DH: **Disentangling the heterogeneity of autism spectrum disorder through genetic findings**. *Nat Rev Neurol* 2014, **10**:74–81.

78. Mannion A, Leader G: **Comorbidity in autism spectrum disorder: A literature review**. *Res Autism Spectr Disord* 2013, **7**:1595–1616.

79. Devlin B, Scherer SW: **Genetic architecture in autism spectrum disorder**. *Curr Opin Genet Dev* 2012, **22**:229–237.

80. Hallmayer J, Cleveland S, Torres A, Phillips J, Cohen B, Torigoe T, Miller J, Fedele A, Collins J, Smith K, Lotspeich L, Croen LA, Ozonoff S, Lajonchere C, Grether JK, Risch N: **Genetic heritability and shared environmental factors among twin pairs with autism.** *Arch Gen*

Psychiatry 2011, **68**:1095–102.

81. Ozonoff S, Young GS, Carter A, Messinger D, Yirmiya N, Zwaigenbaum L, Bryson S, Carver LJ, Constantino JN, Dobkins K, Hutman T, Iverson JM, Landa R, Rogers SJ, Sigman M, Stone WL: **Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study.** *Pediatrics* 2011, **128**:e488–95.

82. Pieretti M, Zhang F, Fu Y-H, Warren ST, Oostra BA, Caskey CT, Nelson DL: **Absence of expression of the FMR-1 gene in fragile X syndrome**. *Cell* 1991, **66**:817–822.

83. Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY: **Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2.** *Nat Genet* 1999, **23**:185–8.

84. Vorstman JAS, Staal WG, van Daalen E, van Engeland H, Hochstenbach PFR, Franke L: **Identification of novel autism candidate regions through analysis of reported cytogenetic abnormalities associated with autism**. *Mol Psychiatry* 2006, **11**:18–28.

85. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D, Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Alvarez VA, Sabatini BL, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Consortium TGO, al. et, Bayés A, Lagemaat LN van de, Collins MO, et al.: **Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses.** *Neuron* 2011, **70**:898–907.

86. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PMA, Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garris M, Soorya L, Klei L, Piven J, Meyer KJ, Anagnostou E, Sakurai T, et al.: **Autism genome-wide copy number variation reveals ubiquitin and neuronal genes**. *Nature* 2009, **459**:569–573.

87. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, Chu SH, Moreau MP, Gupta AR, Thomson SA, Mason CE, Bilguvar K, Celestino-Soper PBS, Choi M, Crawford EL, Davis L, Wright NRD, Dhodapkar RM, DiCola M, DiLullo NM, Fernandez T V, Fielding-Singh V, Fishman DO, Frahm S, Garagaloyan R, Goh GS, Kammela S, Klei L, Lowe JK, Lund SC, et al.: **Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism.** *Neuron* 2011, **70**:863–85.

88. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, McCracken E, Niyazov D, Leppig K, Thiese H, Hummel M, Alexander N, Gorski J, Kussmann J, Shashi V, Johnson K, Rehder C, Ballif BC, Shaffer LG, Eichler EE: **A copy number variation morbidity map of developmental delay.** *Nat Genet* 2011, **43**:838–46.

89. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin C-F, Stevens C, Wang L-S, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, Cai G, Lihm J, Dannenfelser R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu Y, et al.: **Patterns and rates of exonic de novo mutations in autism spectrum disorders.** *Nature* 2012, **485**:242–5.

90. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, Smith JD, Paeper B, Nickerson DA, Dea J, Dong S, Gonzalez LE, Mandell JD, Mane SM, Murtha MT, Sullivan CA, Walker MF, Waqar Z, Wei L, Willsey AJ, Yamrom B, Lee Y, Grabowska E, Dalkic E, Wang Z, Marks S, et al.: **The contribution of de novo coding mutations to autism spectrum disorder**. *Nature* 2014, **515**:216–221.

91. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH: **Missing heritability**

and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010, **11**:446–450.

92. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, Akey JM: **Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants.** *Nature* 2013, **493**:216–20.

93. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG: **Recent and ongoing selection in the human genome.** *Nat Rev Genet* 2007, **8**:857–68.

94. Gibson G: **Rare and common variants: twenty arguments.** *Nat Rev Genet* 2011, **13**:135–45.

95. Pritchard JK: **Are rare variants responsible for susceptibility to complex diseases?** *Am J Hum Genet* 2001, **69**:124–37.

96. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, Hardison M, Person R, Bekheirnia MR, Leduc MS, Kirby A, Pham P, Scull J, Wang M, Ding Y, Plon SE, Lupski JR, Beaudet AL, Gibbs RA, Eng CM: **Clinical whole-exome sequencing for the diagnosis of mendelian disorders.** *N Engl J Med* 2013, **369**:1502–11.

97. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K, Munson J, Hiatt JB, Turner EH, Levy R, O'Day DR, Krumm N, Coe BP, Martin BK, Borenstein E, Nickerson DA, Mefford HC, Doherty D, Akey JM, Bernier R, Eichler EE, Shendure J, Kryukov G V., Shpunt A, Stamatoyannopoulos JA, Sunyaev SR, et al.: **Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders.** *Science* 2012, **338**:1619–22.

98. Biesecker LG, Spinner NB: **A genomic view of mosaicism and human disease**. *Nat Rev Genet* 2013, **14**:307–320.

99. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M: **Genome-wide association studies in diverse populations.** *Nat Rev Genet* 2010, **11**:356–66.

100. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C, Sanger F, al. et, Maxam AM, Gilbert W, Consortium IHGS, Schloss JA, Margulies M, al. et, Valouev A, al. et, Metzker ML, Liu L, al. et, Ju J, al. et, Shendure J, al. et, Pushkarev D, al. et, Schadt EE, al. et, Eid J, al. et, Wang Z, al. et, Park PJ, et al.: **Ten years of next-generation sequencing technology.** *Trends Genet* 2014, **30**:418–26.

101. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions**. *Nat Rev Genet* 2011, **13**:36.

102. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornson RD, Carriero NJ, Meyer KA, Bilguvar K, Mane SM, Šestan N, Lifton RP, Günel M, Roeder K, Geschwind DH, Devlin B, State MW: **De novo mutations revealed by whole-exome sequencing are strongly associated with autism**. *Nature* 2012, **485**:237–241.

103. Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, Epstein MP, Glauser T, Goldstein DB, Han Y, Heinzen EL, Hitomi Y, Howell KB, Johnson MR, Kuzniecky R, Lowenstein DH, Lu Y-F, Madou MRZ, Marson AG, Mefford HC, Esmaeeli Nieh S, O'Brien TJ, Ottman R, Petrovski S, Poduri A, Ruzzo EK, Scheffer IE, Sherr EH, Yuskaitis CJ, Abou-Khalil B, et al.: **De novo mutations in epileptic encephalopathies.** *Nature* 2013, **501**:217–21.

104. Rideout WM, Coetzee GA, Olumi AF, Jones PA: **5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes.** *Science* 1990, **249**:1288–90.

105. Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, Keinan A, Siepel A: **Genome-wide**

inference of natural selection on human transcription factor binding sites. *Nat Genet* 2013, **45**:723–9.

106. Blake RD, Hess ST, Nicholson-Tuell J: **The influence of nearest neighbors on the rate and pattern of spontaneous point mutations.** *J Mol Evol* 1992, **34**:189–200.

107. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kähäri AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, et al.: **Ensembl 2014.** *Nucleic Acids Res* 2014, **42**(Database issue):D749–55.

108. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM: **RefSeq: an update on mammalian reference sequences.** *Nucleic Acids Res* 2014, **42**(Database issue):D756–63.

109. The International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437**:1299–320.

110. Okae H, Chiba H, Hiura H, Hamada H, Sato A, Utsunomiya T, Kikuchi H, Yoshida H, Tanaka A, Suyama M, Arima T: **Genome-wide analysis of DNA methylation dynamics during early human development.** *PLoS Genet* 2014, **10**:e1004868.

111. Hovestadt V, Jones DTW, Picelli S, Wang W, Kool M, Northcott PA, Sultan M, Stachurski K, Ryzhova M, Warnatz H-J, Ralser M, Brun S, Bunt J, Jäger N, Kleinheinz K, Erkek S, Weber UD, Bartholomae CC, von Kalle C, Lawerenz C, Eils J, Koster J, Versteeg R, Milde T, Witt O, Schmidt S, Wolf S, Pietsch T, Rutkowski S, Scheurlen W, et al.: **Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing.** *Nature* 2014, **510**:537–41.

112. Campbell MC, Tishkoff SA: **African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping.** *Annu Rev Genomics Hum Genet* 2008, **9**:403–33.

113. Schaffner SF: **The X chromosome in population genetics.** *Nat Rev Genet* 2004, **5**:43–51.

114. Mugal CF, Ellegren H: **Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content.** *Genome Biol* 2011, **12**:R58.

115. Walser J-C, Furano A V: **The mutational spectrum of non-CpG DNA varies with CpG content.** *Genome Res* 2010, **20**:875–82.

116. Kamiya H, Tsuchiya H, Karino N, Ueno Y, Matsuda A, Harashima H: **Mutagenicity of 5-formylcytosine, an oxidation product of 5-methylcytosine, in DNA in mammalian cells.** *J Biochem* 2002, **132**:551–5.

117. Deaton AM, Bird A: **CpG islands and the regulation of transcription.** *Genes Dev* 2011, **25**:1010–22.

118. Levinson G, Gutman GA: **Slipped-strand mispairing: a major mechanism for DNA sequence evolution.** *Mol Biol Evol* 1987, **4**:203–21.

119. Panchin AY, Mitrofanov SI, Alexeevski A V, Spirin SA, Panchin Y V: **New words in human mutagenesis.** *BMC Bioinformatics* 2011, **12**:268.

120. Lanfear R, Welch JJ, Bromham L: **Watching the clock: studying variation in rates of molecular evolution between species.** *Trends Ecol Evol* 2010, **25**:495–503.

121. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark

AG: **Natural selection on protein-coding genes in the human genome.** *Nature* 2005, **437**:1153–7.

122. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, et al.: **Genome-wide detection and characterization of positive selection in human populations.** *Nature* 2007, **449**:913–8.

123. Alföldi J, Lindblad-Toh K: **Comparative genomics as a tool to understand evolution and disease.** *Genome Res* 2013, **23**:1063–8.

124. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7**:248–9.

125. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB: **Genic intolerance to functional variation and the interpretation of personal genomes.** *PLoS Genet* 2013, **9**:e1003709.

126. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nat Genet* 2014, **46**:310–315.

127. Georgi B, Voight BF, Bućan M: **From mouse to human: evolutionary genomics analysis of human orthologs of essential genes.** *PLoS Genet* 2013, **9**:e1003484.

128. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4**:44–57.

129. Hamdan FF, Srour M, Capo-Chichi J-M, Daoud H, Nassif C, Patry L, Massicotte C, Ambalavanan A, Spiegelman D, Diallo O, Henrion E, Dionne-Laporte A, Fougerat A, Pshezhetsky A V., Venkateswaran S, Rouleau GA, Michaud JL: **De Novo Mutations in Moderate or Severe Intellectual Disability**. *PLoS Genet* 2014, **10**:e1004772.

130. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N, Dufke A, Cremer K, Hempel M, Horn D, Hoyer J, Joset P, Röpke A, Moog U, Riess A, Thiel CT, Tzschach A, Wiesener A, Wohlleber E, Zweier C, Ekici AB, Zink AM, Rump A, Meisinger C, Grallert H, Sticht H, et al.: **Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study.** *Lancet* 2012, **380**:1674–82.

131. **Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability** [http://www.nejm.org/doi/full/10.1056/NEJMoa1206524]

132. **Large-scale discovery of novel genetic causes of developmental disorders.** *Nature* 2015, **519**:223–8.

133. Cooper GM, Shendure J: **Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.** *Nat Rev Genet* 2011, **12**:628–40.

134. Uddin M, Tammimies K, Pellecchia G, Alipanahi B, Hu P, Wang Z, Pinto D, Lau L, Nalpathamkalam T, Marshall CR, Blencowe BJ, Frey BJ, Merico D, Yuen RKC, Scherer SW: **Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder.** *Nat Genet* 2014, **46**:742–7.

135. Ginsburg D, Bowie EJ: **Molecular genetics of von Willebrand disease.** *Blood* 1992, **79**:2507–19.

136. Orosco LA, Ross AP, Cates SL, Scott SE, Wu D, Sohn J, Pleasure D, Pleasure SJ, Adamopoulos IE, Zarbalis KS: **Loss of Wdfy3 in mice alters cerebral cortical neurogenesis**

reflecting aspects of the autism pathology. *Nat Commun* 2014, **5**:4692.

137. Gallie BL, Campbell C, Devlin H, Duckett A, Squire JA: **Developmental basis of retinal-specific induction of cancer by RB mutation.** *Cancer Res* 1999, **59**(7 Suppl):1731s–1735s.

138. Vogel F, Rathenberg R: **Spontaneous Mutation in Man**. In *Advances in Human Genetics*. Boston, MA: Springer US; 1975:223–318.

139. Mancini D, Singh S, Ainsworth P, Rodenhiser D: **Constitutively methylated CpG dinucleotides as mutation hot spots in the retinoblastoma gene (RB1).** *Am J Hum Genet* 1997, **61**:80–7.

140. Richter S, Vandezande K, Chen N, Zhang K, Sutherland J, Anderson J, Han L, Panton R, Branco P, Gallie B: **Sensitive and Efficient Detection of RB1 Gene Mutations Enhances Care for Families with Retinoblastoma**. *Am J Hum Genet* 2003, **72**:253–269.

141. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R: **Ascertainment bias in studies of human genome-wide polymorphism.** *Genome Res* 2005, **15**:1496–502.

142. Cooper DN, Youssoufian H: **The CpG dinucleotide and human genetic disease**. *Hum Genet* 1988, **78**:151–155.

143. Mort M, Ivanov D, Cooper DN, Chuzhanova NA: **A meta-analysis of nonsense mutations causing human genetic disease.** *Hum Mutat* 2008, **29**:1037–47.

144. Hogg A, Bia B, Onadim Z, Cowell JK: **Molecular mechanisms of oncogenic mutations in tumors from patients with bilateral and unilateral retinoblastoma.** *Proc Natl Acad Sci U S A* 1993, **90**:7351–5.

145. Zhang K, Nowak I, Rushlow D, Gallie BL, Lohmann DR: **Patterns of missplicing caused by RB1 gene mutations in patients with retinoblastoma and association with phenotypic expression.** *Hum Mutat* 2008, **29**:475–84.

146. Lohmann D, Brandt B, Hopping W, Passarge E, Horsthemke B: **Distinct RB1 gene mutations with low penetrance in hereditary retinoblastoma**. *Hum Genet* 1994, **94**.

147. Lee JO, Russo AA, Pavletich NP: **Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7.** *Nature* 1998, **391**:859–65.

148. Sánchez-Sánchez F, Ramírez-Castillejo C, Weekes DB, Beneyto M, Prieto F, Nájera C, Mittnacht S: **Attenuation of disease phenotype through alternative translation initiation in low-penetrance retinoblastoma**. *Hum Mutat* 2007, **28**:159–167.

149. Consortium EA, Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, O'Donnell-Luria A, Ware J, Hill A, Cummings B, Tukiainen T, Birnbaum D, Kosmicki J, Duncan L, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Cooper D, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki M, et al.: *Analysis of Protein-Coding Genetic Variation in 60,706 Humans*. Cold Spring Harbor Labs Journals; 2015.

150. Brogna S, Wen J: **Nonsense-mediated mRNA decay (NMD) mechanisms.** *Nat Struct Mol Biol* 2009, **16**:107–13.

151. Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT: **LOVD v.2.0: the next generation in gene variant databases**. *Hum Mutat* 2011, **32**:557–563.

152. Dudley JT, Kim Y, Liu L, Markov GJ, Gerold K, Chen R, Butte AJ, Kumar S: **Human genomic disease variants: a neutral evolutionary explanation.** *Genome Res* 2012, **22**:1383–94.

153. McVicker G, Gordon D, Davis C, Green P: **Widespread genomic signatures of natural selection in hominid evolution.** *PLoS Genet* 2009, **5**:e1000471.

154. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ: **Testing for an unusual distribution of rare variants.** *PLoS Genet* 2011, **7**:e1001322.

155. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD, State MW, Devlin B, Roeder K, Sanders S, Murtha M, Gupta A, Murdoch J, Raubeson M, Neale B, Kou Y, Liu L, Ma'ayan A, Samocha K, O'Roak B, Vives L, Girirajan S, Karakoc E, Krumm N, Iossifov I, Ronemus M, et al.: **Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes**. *PLoS Genet* 2013, **9**:e1003671.

156. Ware JS, Samocha KE, Homsy J, Daly MJ: **Interpreting de novo Variation in Human Disease Using denovolyzeR.** *Curr Protoc Hum Genet* 2015, **87**:7.25.1–15.

157. Siepel A, Haussler D: **Phylogenetic estimation of context-dependent substitution rates by maximum likelihood.** *Mol Biol Evol* 2004, **21**:468–88.

158. Excoffier L, Yang Z: **Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees.** *Mol Biol Evol* 1999, **16**:1357–68.

159. Wolfe KH, Sharp PM, Li W-H: **Mutation rates differ among regions of the mammalian genome**. *Nature* 1989, **337**:283–285.

160. Subramanian S, Kumar S: **Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes.** *Genome Res* 2003, **13**:838–44.

161. Kalari KR, Casavant M, Bair TB, Keen HL, Comeron JM, Casavant TL, Scheetz TE: **First exons and introns--a survey of GC content and gene structure in the human genome.** *In Silico Biol* 2006, **6**:237–42.

162. Burns MB, Temiz NA, Harris RS: **Evidence for APOBEC3B mutagenesis in multiple human cancers**. *Nat Genet* 2013, **45**:977–983.

163. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov G V, Carter SL, Saksena G, Harris S, Shah RR, Resnick MA, Getz G, Gordenin DA: **An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers**. *Nat Genet* 2013, **45**:970–976.

164. Zhu YO, Siegal ML, Hall DW, Petrov DA: **Precise estimates of mutation rate and spectrum in yeast**. *Proc Natl Acad Sci* 2014, **111**:E2310–E2318.

165. Eyre-Walker A, Eyre-Walker YC: **How much of the variation in the mutation rate along the human genome can be explained?** *G3 (Bethesda)* 2014, **4**:1667–70.

166. Kimura M, Ohta T: **On some principles governing molecular evolution.** *Proc Natl Acad Sci U S A* 1974, **71**:2848–52.

167. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, et al.: **Great ape genetic diversity and population history**. *Nature* 2013, **499**:471–475.

168. Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, Hu Y, He W, Zhang S, Fan W, Zhu L, Li D, Zhang X, Chen Q, Zhang H, Zhang Z, Jin X, Zhang J, Yang H, Wang JJ, Wang JJ, Wei F: **Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation.** *Nat Genet* 2013, **45**:67–71.

169. Hussin JG, Hodgkinson A, Idaghdour Y, Grenier J-C, Goulet J-P, Gbeha E, Hip-Ki E, Awadalla P: **Recombination affects accumulation of damaging and disease-associated**

mutations in human populations. *Nat Genet* 2015, **47**:400–404.

170. Koren A, Handsaker RE, Kamitaki N, Karlić R, Ghosh S, Polak P, Eggan K, McCarroll SA: **Genetic Variation in Human DNA Replication Timing**. *Cell* 2014, **159**:1015–1026.

171. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB: **Genic intolerance to functional variation and the interpretation of personal genomes.** *PLoS Genet* 2013, **9**:e1003709.

172. Keightley PD: **Rates and fitness consequences of new mutations in humans.** *Genetics* 2012, **190**:295–304.

173. Sniegowski PD, Gerrish PJ, Lenski RE: **Evolution of high mutation rates in experimental populations of E. coli**. *Nature* 1997, **387**:703–705.

174. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, Georgieva L, Rees E, Palta P, Ruderfer DM, Carrera N, Humphreys I, Johnson JS, Roussos P, Barker DD, Banks E, Milanova V, Grant SG, Hannon E, Rose SA, Chambert K, Mahajan M, Scolnick EM, Moran JL, Kirov G, Palotie A, McCarroll SA, Holmans P, Sklar P, Owen MJ, et al.: **De novo mutations in schizophrenia implicate synaptic networks.** *Nature* 2014, **506**:179–84.